

# From Insights to Impact

## Actionable Interpretability for Neural Machine Translation

Gabriele Sarti

# **From Insights to Impact: Actionable Interpretability for Neural Machine Translation**

Gabriele Sarti



/ university of  
groningen

**CLCG**  
Center for Language and Cognition Groningen



The research reported in this thesis has been carried out under the auspices of the Center for Language and Cognition Groningen (CLCG) of the Faculty of Arts of the University of Groningen.

The project was funded by the Dutch Research Council (NWO), under the NWA-ORC project *InDeep: Interpreting Deep Learning Models for Text and Sound* (NWA.1292.19.399)



Groningen Dissertations in Linguistics 263

© 2025, Gabriele Sarti

Document typeset with Quarto and compiled using pdfLaTeX

Cover design: Gabriele Sarti and Lukas Edman



university of  
groningen

# From Insights to Impact: Actionable Interpretability for Neural Machine Translation

## PhD Thesis

to obtain the degree of PhD at the  
University of Groningen  
on the authority of the  
Rector Magnificus Prof. J. M. A. Scherpen  
and in accordance with  
the decision by the College of Deans.

The thesis will be defended in public on

Thursday 11 December 2025 at 14:30 hours

by

**Gabriele Sarti**

born on 11 November 1996  
in Trieste, Italy

**Supervisors**

Dr. A. Bisazza

Prof. M. Nissim

**Co-supervisor**

Dr. G. Chrupała

**Assessment Committee**

Prof. M. Biehl

Prof. B. Plank

Prof. I. Titov

*A nonno Giancarlo, e alle sue simmetrie nell'arte e nella scienza*



# Acknowledgements

Outsiders to the academic world might see the PhD as a lonely, long-winding path in the pursuit of knowledge. On the contrary, looking back over the past four years, it's clear to me that it truly takes a village to raise happy, fulfilled PhD students. The following few pages pay homage to my own little village, to which I owe the wonderful experience that is now reaching its conclusion.

This thesis would not have been possible without the thoughtful guidance of my supervisors, Arianna Bisazza, Malvina Nissim and Grzegorz Chrupała. Arianna, I cannot thank you enough for giving me a chance to embark on this journey in Groningen. Your outstanding curiosity and passion for research were a constant source of motivation for me to keep asking difficult questions, dig deeper into promising directions, and never stop learning new things. I cherish the time we spent together, at the office debating promising new directions and asking ourselves “What does it take to make interpretability useful in practice?”, playing Codenames at lab retreats, or enjoying amazing meals at conferences all around the world. Your keen eye for designing and executing compelling research agendas and your genuine attention to the well-being of everyone in the lab have been great sources of inspiration for the kind of scientist I strive to become. Malvina, I cannot imagine my PhD journey without your unwavering support. Your cheerfulness and excitement have always made challenging moments feel lighter, and I thank you for being the voice of reason that often helped me strike a balance between work and life matters. Working with you taught me that research is, in large part, about *community*, and that the time invested in sustaining and growing the communities we genuinely care about is time well spent. Your commitment to making advances in our field accessible to the general public sparked my interest in science communication, and your work with various communities to ensure that your research can be inclusive and useful motivated me to adopt a user-centric perspective in my interpretability work. I hope our paths will cross again, especially if it's for a chat over a cecina in Pisa, or at the next CLiC-it! Grzegorz, I am grateful that you agreed to join us as my co-supervisor for this journey. Thank you for consistently contributing interesting perspectives during our InDeep meetings and for your invaluable support

for our PECoRE and QE4PE projects. I am still looking forward to tasting one of your famed espressos!

I am also thankful to the members of my reading committee, Profs. Barbara Plank, Ivan Titov and Michael Biel, for agreeing to spend some time on this (admittedly lengthy) manuscript and providing me with valuable feedback to improve it. Also, thank you to Profs. Eva Vanmassenhove, Rik van Noord and Jelle Zuidema for joining my defense committee and bringing in your expertise on machine translation, bias and interpretability.

My research project was funded and carried out as part of the NWO InDeep consortium, led by Prof. Jelle Zuidema. A big thank you to all professors — Afra Alishahi, Anstke Fokkens, Ashley Burgoyné, Iris Hendrickx, Louis ten Bosch, Martijn Bentum, Tom Lenz — and to my fellow PhDeeps — Charlotte Pouw, Gaofei Shen, Hosein Mohebbi, Marcel Vélez, Jane Arleth dela Cruz, Jonathan Kamp — for always making me feel part of a thriving research community working together to make deep learning models more transparent for applications in speech, text and music. Thank you also to the honorary UvA PhDeeps — Michael Hanna, Marianne de Heer Kloots, Oskar van der Wal — who were always involved in our meetings, and to Rudmer Canjels for handling the consortium activities and funding our (occasional) virgin mojito.

The GroNLP group at the University of Groningen has been my research home for the past four years, and remains one of the most friendly and welcoming work environments I have seen to this day. A big thank you to all my wonderful colleagues, past and present — Gertjan van Noord, Johan Bos, Martijn Wieling, Gosse Bouma, Andreas van Cranenburgh, Antonio Toral, Federico Pianzola, Khalid Al Khatib, Defne Abur, Frank Tsiwah, Lisa Bylinina, Kun He, Huiyuan Lai, Wietse de Vries, Martijn Bartelds, Chunliu Wang, Teja Rebernik, Prajit Dhar, Masha Medvedeva, Raoul Buurke, Silvia Stopponi, Thomas Tienkamp, Iris Schepers, Hedwig Sekeres, Xiao Zhang, Greta Zella, Reihaneh Amouie, Noa Visser, Katharina Polsterer, Xiaoyan Yang, Yun Hao, Ze Yu, Yanan Wu, Yongjian Chen, Nastja Shaboltas, Kyo Gerrits, Xiaolu Wang, Susan Lotz, Shaozhen Shi, Sijbren van Vaals, Sara Nabhani, Dan Mu, Ruhi Mahadeshwar, Valentine Lucquiault, Ezgi Başar and Gijs Danoe. A special thanks to fellow members of the InCLow team — Jaap Jumelet, Jirui Qi, Francesca Padovani, Yuqing Zhang and Yevgen Matusevych — with whom I had a chance to share our cozy office, discuss interesting work at our reading groups, or spend some quality time together in Suzhou. Francesca and many other colleagues — Pritha Majumdar, Leonidas Zotos, Franziska Pannach and Maria Pilar Uribe Silva — joined the group after me, bringing in a fresh breath of air and enthusiasm for spending time together in Groningen. Thank you

all for having filled my days in Groningen with nice chats and get-togethers! Ahmet Üstün, I'm happy we stayed in touch after the end of your PhD and for the interesting discussions, ranging from hardcore technical topics to Mutehşem Yüzil over a coffee or a tantuni!

In the many publications that filled the past few years, I even had the chance to collaborate with some colleagues towards interesting research ideas — Lukas Edman, Ana Guerberof, Tommaso Caselli and Daniel Scalena. Joachim Lukas, while we were sharing a room in Abu Dhabi you revealed your bropensity for joking around, and since then our (mostly non-research) conversations have always been a wild ride. Looking forward to many years of hanging out at conferences around the world! Ana, thank you for always being a cheerful presence in the office and for always bringing your no-nonsense attitude to keep our project meetings grounded in reality. Your expertise on post-editing and translation workflows has been essential to ensuring the soundness of user studies in this thesis. Tommaso, I enjoyed our collaboration on the rebus paper, but I enjoyed the spaghettiate partigiane at your place on April 25th even more! Looking forward to collaborating on making attribution useful for socially relevant applications. Last but not least, Daniel, I am grateful to have had the chance to collaborate with you on so many interesting projects and to share lovely memories in Singapore, Miami and Boston. Seeing you grow into a talented researcher has been very inspiring, and I'll always be on the lookout for your next outstanding interpretability paper!

Finally, the GroNLP acknowledgements would not be complete without thanking the many people that left their mark in our group with their visits — Arianna Graciotti, Arianna Muti, Daniela Occhipinti, Beatrice Savoldi, Serena Coschignano, Giulia Rambelli, Elena Sofia Ruzzetti, Janica Hackenbuchner, Anaís Almendra, Akari Haga, Saad Amin, Okky Ibromhim — and all the students whom I had the chance to meet and supervise as part of the Advanced NLP course and the LCT masters' program, in particular Ludwig Sickert, Qiankun Zheng, Konstantin Chernyshev and Khondoker Islam. I also acknowledge the support of the Center of Information Technology at the University of Groningen, which supported my research by providing free access to the Peregrine and Habrók clusters, and the SURF cooperative, which granted me free access to their Snellius supercomputer. Finally, I should also thank Deutsche Bahn and Arriva for making every one of my weekly trips between Groningen and Bremen an adventure, and for providing me with a lot of time to ponder over my research questions after missing connections in Leer. I hope my great-grandchildren will manage to see the mythical Wunderline reaching completion in their lifetime.

Aside from our group in Groningen, I am fortunate to have established connections with

many outstanding researchers worldwide. First, a warm thank you to the colleagues at the ItaliaNLP Lab — Felice dell’Orletta, Alessio Miaschi, Chiara Alzetta, Giulia Venturi, Dominique Brunato and Cristiano Ciaccio — who introduced me to NLP research in the year 2019 BCE (Before the ChatGPT Era), and who remain to this day close collaborators for Italian NLP projects. I am also thankful to the Amazon Translate team, and in particular to Georgiana Dinu, Maria Nadejde and Xing Niu for welcoming me into their team during my first year of PhD and for helping me discover the exciting world of industrial research. In the past few years, I have been privileged to collaborate with outstanding fellow PhDs — Nils Feldhus, Hosein Mohebbi, Javier Ferrando and Vilém Zouhar. Nils, Hosein, Javier and Vilém, our discussions and your thoughtful support laid the foundation for some of my most impactful works to date. You are all incredibly talented scientists and I hope to have the chance to work together again someday! Alongside Hosein, I also want to thank all the other co-organizers of BlackboxNLP 2025 — Yonatan Belinkov, Aaron Mueller, Najoung Kim, Hanjie Chen, Dana Arad and Martin Tutek — for giving me a chance to be involved in the organization of this year’s edition of the workshop, and for working together efficiently to make it a success. A special shoutout to Dana, whose in-person problem-solving skills were fundamental to ensuring everything ran smoothly on the day of the event. I want to also sincerely thank Luca Bortolussi — who always kept me in the loop and included me in the initiatives and activities of the Trieste AI community — and the many members of the UniTS AILab — in particular Emanuele Ballarin, Sara Candussio, Davide Scassola, Gaia Saveri, Francesca Cairoli, Francesco Giacomarra — for always making me feel welcome in my moments back in Trieste. Finally, I want to thank David Bau and the members of the BauLab for their warm welcome this summer at NEMI. David, you are an inspiring researcher (coincidentally, the cover of this thesis was inspired by your 3D-printed aperiodic monotiles!), and I thank you for sharing my vision for an open interpretability ecosystem that benefits everyone. I am excited to join you soon in Boston and to keep advancing the frontier of interpretability research together.

Beyond work, I want to thank everyone else who made these years unforgettable. First, friends in Germany — Sergei Kolesnikov, Anastasiia Nikitina, Gleb Tomachevski, Perancha Domingo, Victor Martinez, Priscilla Oh — who made me feel right at home during my commuter life between Groningen and Bremen. Sergei, thank you for being an excellent teacher and getting me from zero to hero in Russian in no time! Our dinners and board game evenings with Lera and Nastia, and our soccer matches with Gleb are among my happiest memories of these years. Thanks also to all my Italian friends all around the world — Mario Julianelli, Giuseppe Attanasio, Andrea Santilli, Francesco

Cicala, Salvatore Milite, Andrea Gasparin, Leonardo Stincone, Andrea Lorenzon — I am glad we are still in touch, to discuss exciting new projects, or simply to update each other about our lives. Thank you also to my long-time friends from Galilei — Ethan Turco, Elia Miraz, Jimmi d'Ambrosi, Enrico Pieri, Lorenzo Macor, Simone Ramazzotti — with whom I still share interesting conversations over a Negroni pitcher after more than ten years apart.

This section would not be complete without thanking some very special people. First, my amazing paranympths, Otje Minnema and Sara Gemelli, who have been two of my closest friends throughout my PhD path, and an essential help in preparing for my defense. Your presence kept me grounded when all the stress and the traveling were taking a toll, and you made Groningen feel like home from the very beginning of this journey — I loved our time together in the Noorderplantsoen, our trips to Schier, and our dinners on the Pannekoekschip. It has been a real pleasure to work together towards our "tesi ostiche, contorte e sconcertanti". Lunga vita al triumphonato! Mattia, despite our physical distance and our busy schedules, I never felt distant from you for a second in all these years, knowing that at the next meeting we would both feel right at home with each other, as if we never left Trieste. Thank you for being a steady presence in my life, and for teaching me every day what friendship is really about.

I want to thank my family for always being there to cheer for my successes and comfort me in challenging times from Trieste, Terni and Rome. Dad, mom, Barbara and Diego — your love is the solid foundation at the base of all my successes, and your unwavering support has given me, time and time again, the confidence to leave my safe harbor and set sail towards unknown waters. Words fail me in expressing how much this matters to me.

Dulcis in fundo, I want to express my heartfelt appreciation to my partner, Valeriya, who has been my constant through every variable and my solution for every problem throughout this journey. Your ever-loving presence and help made even the most challenging moments feel easy. Thank you for being my steady ground, my voice of reason and the love of my life.



# Contents

<b>Contents</b>	<b>xiii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Outline and Contributions . . . . .	3
1.2 Scientific Output . . . . .	6
1.2.1 Main Publications . . . . .	6
1.2.2 Open-source Contributions . . . . .	8
1.2.3 Other Research Contributions . . . . .	9
<b>2 Background</b>	<b>13</b>
2.1 From Neural Networks to Neural Language Models . . . . .	14
2.1.1 Supervised Learning for Neural Networks . . . . .	14
2.1.2 Transformer Neural Networks . . . . .	15
2.1.3 Transformer Language Models . . . . .	18
2.2 Explaining Predictions with Input Attribution . . . . .	21
2.2.1 Attribution Method Categories . . . . .	22
2.2.2 Evaluating and Using Attribution Methods . . . . .	25
2.3 Conditioning Language Model Generations . . . . .	26
2.3.1 Controlling Input Context . . . . .	27
2.3.2 Controlling Model Representations . . . . .	28
2.4 Machine Translation . . . . .	31
2.5 MT Post-Editing and Evaluation . . . . .	34
2.5.1 Post-editing MT . . . . .	34
2.5.2 MT Evaluation . . . . .	36
2.6 Quality Estimation for MT . . . . .	38
<b>I Attributing Context Usage in Multilingual NLP</b>	<b>43</b>
<b>3 Attributing Language Model Generations with the Inseq Toolkit</b>	<b>45</b>
3.1 Related Work . . . . .	47
3.2 Design . . . . .	48
3.2.1 Guiding Principles . . . . .	49
3.2.2 Input Attribution and Post-processing . . . . .	50
3.2.3 Customizing generation and attribution . . . . .	51
3.2.4 Usability Features . . . . .	52
3.3 Case Studies . . . . .	54

3.3.1	Gender Bias in Machine Translation . . . . .	54
3.3.2	Locating Factual Knowledge inside GPT-2 . . . . .	56
3.4	Conclusion . . . . .	58
<b>4</b>	<b>Quantifying Context Usage in Neural Machine Translation</b>	<b>59</b>
4.1	Introduction . . . . .	60
4.2	Related Work . . . . .	62
4.3	The PECoRE Framework . . . . .	63
4.3.1	Notation . . . . .	64
4.3.2	Context-sensitive Token Identification (CTI) . . . . .	65
4.3.3	Contextual Cues Imputation (CCI) . . . . .	65
4.4	Context Reliance Plausibility in Context-aware MT . . . . .	69
4.4.1	Experimental Setup . . . . .	69
4.4.2	Metrics for Context-sensitive Target Identification . . . . .	72
4.4.3	Plausibility Evaluation Metrics . . . . .	73
4.4.4	CTI Plausibility Results . . . . .	74
4.4.5	Methods for Contextual Cues Imputation . . . . .	75
4.4.6	CCI Plausibility Results . . . . .	76
4.5	Detecting Context Reliance in the Wild . . . . .	77
4.6	Integrating PECoRE in Inseq . . . . .	79
4.7	Conclusion . . . . .	81
<b>5</b>	<b>Answer Attribution for Trustworthy Retrieval-Augmented Generation</b>	<b>83</b>
5.1	Introduction . . . . .	84
5.2	Background and Related Work . . . . .	86
5.2.1	Answer Attribution Methods . . . . .	86
5.3	Method . . . . .	88
5.3.1	From Granular Attributions to Document-level Citations . . . . .	90
5.4	Agreement with Human Answer Attribution Annotations . . . . .	91
5.4.1	Experimental Setup . . . . .	91
5.4.2	Entailment-based Baselines . . . . .	92
5.4.3	Results and Analysis . . . . .	93
5.5	Answer Attribution for Long-form QA . . . . .	94
5.5.1	Experimental Setup . . . . .	95
5.5.2	Results . . . . .	96
5.5.3	Qualitative Analysis of Disagreements . . . . .	97
5.6	Limitations . . . . .	100

---

5.7	Conclusion . . . . .	101
<b>II</b>	<b>Conditioning Generation for Personalized Machine Translation</b>	<b>103</b>
<b>6</b>	<b>Retrieval and Marking for Attribute-Controlled Translation</b>	<b>105</b>
6.1	Introduction . . . . .	106
6.2	Method . . . . .	108
6.2.1	Our Approach: RAMP . . . . .	108
6.2.2	Cross-Lingual Prompting . . . . .	109
6.3	Experiments . . . . .	109
6.3.1	Datasets . . . . .	110
6.3.2	Large Language Models . . . . .	110
6.3.3	Baseline . . . . .	111
6.3.4	Evaluation Metrics . . . . .	111
6.3.5	Same-Language Prompting . . . . .	112
6.3.6	Cross-Lingual Prompting . . . . .	114
6.4	Limitations . . . . .	115
6.5	Conclusion . . . . .	116
<b>7</b>	<b>Steering Language Models for Personalized Machine Translation</b>	<b>117</b>
7.1	Introduction . . . . .	118
7.2	Related Work . . . . .	120
7.3	Preliminaries . . . . .	120
7.3.1	Are Personalized Translations Discernible? . . . . .	122
7.3.2	Can LLMs Reproduce Human Translation Styles? . . . . .	123
7.3.3	Finding Personalization Information in LLM Representations .	124
7.4	Methods . . . . .	125
7.4.1	Prompting Baselines . . . . .	125
7.4.2	Steering Baselines . . . . .	126
7.4.3	Contrastive SAE Steering . . . . .	126
7.5	Experiments . . . . .	128
7.5.1	Setup . . . . .	128
7.5.2	Results and Discussion . . . . .	131
7.6	Limitations . . . . .	133
7.7	Conclusion . . . . .	134

<b>III Interpretability in Human Translation Workflows</b>	<b>137</b>
<b>8 Machine Translation Post-editing for Typologically Diverse Languages</b>	<b>139</b>
8.1 Introduction . . . . .	140
8.2 Related Work . . . . .	141
8.3 The DivEMT Dataset . . . . .	142
8.3.1 Subjects and Task Scheduling . . . . .	142
8.3.2 Choice of Source Texts . . . . .	144
8.3.3 Choice of Languages . . . . .	144
8.3.4 Choice of MT Systems . . . . .	145
8.3.5 Translation Platform and Collected Data . . . . .	146
8.4 Post-Editing Effort Across Languages . . . . .	147
8.4.1 Temporal Effort and Productivity Gains . . . . .	147
8.4.2 Post-Editing Rate . . . . .	152
8.4.3 Perception of Productivity Gain . . . . .	154
8.5 Limitations . . . . .	156
8.6 Conclusions . . . . .	156
<b>9 Word-level Quality Estimation for Machine Translation Post-editing</b>	<b>159</b>
9.1 Introduction . . . . .	160
9.2 Experimental Setup . . . . .	162
9.2.1 Structure of the Study . . . . .	162
9.2.2 Highlight Modalities . . . . .	163
9.2.3 Data and MT model . . . . .	166
9.2.4 Participants . . . . .	169
9.2.5 Editing Interface . . . . .	169
9.3 Analysis . . . . .	170
9.3.1 Productivity . . . . .	170
9.3.2 Highlights and Edits . . . . .	172
9.3.3 Quality Assessment . . . . .	175
9.3.4 Usability . . . . .	179
9.4 Limitations . . . . .	180
9.5 Broader Impact and Ethical Considerations . . . . .	181
9.6 Conclusion . . . . .	182
<b>10 Unsupervised MT Error Detection and Human Disagreement</b>	<b>183</b>
10.1 Introduction . . . . .	184

10.2 Related Work . . . . .	185
10.3 Models and Datasets . . . . .	186
10.4 Evaluated Metrics . . . . .	188
10.5 Experiments . . . . .	189
10.5.1 Setup . . . . .	189
10.5.2 Results . . . . .	192
10.6 Limitations . . . . .	194
10.7 Conclusion . . . . .	196
<b>11 Conclusion</b>	<b>197</b>
11.1 Research Questions Revisited . . . . .	198
11.2 Outlook and Future Directions . . . . .	201
<b>Bibliography</b>	<b>205</b>
<b>Appendices</b>	<b>259</b>
<b>A Attributing Context Usage in Multilingual NLP</b>	<b>259</b>
A.1 Attributing Language Model Generations with the Inseq Toolkit . . . . .	259
A.1.1 Additional Details on Turkish Gender Bias Case Study . . . . .	259
A.1.2 Example of Pair Aggregation for Contrastive MT Comparison	259
A.1.3 Example of Quantized CAT Attribution . . . . .	259
A.2 Quantifying Context Usage in Neural Machine Translation . . . . .	264
A.2.1 Details on Translation Evaluation . . . . .	264
A.2.2 Full CTI and CCI Results . . . . .	264
A.2.3 PECoRE for Other Language Generation Tasks . . . . .	264
A.3 Answer Attribution for Trustworthy Retrieval-Augmented Generation	271
A.3.1 Answer Attribution on the Full XOR-AttriQA . . . . .	271
A.3.2 ELI5 Evaluation with Standard Prompt . . . . .	271
A.3.3 More Examples of Disagreement . . . . .	273
<b>B Conditioning Generation for Personalized Machine Translation</b>	<b>277</b>
B.1 Retrieval and Marking for Attribute-Controlled Translation . . . . .	277
B.1.1 Prompt Templates . . . . .	277
B.1.2 Full Per-language Results . . . . .	277
B.1.3 Error Analysis of Cross-Lingual Prompting . . . . .	280
B.2 Steering Language Models for Machine Translation Personalization .	282

B.2.1	Experiments Reproducibility . . . . .	282
B.2.2	All Models Results . . . . .	284
B.2.3	Dataset and Generation Examples . . . . .	284
<b>C</b>	<b>Interpretability in Human Translation Workflows</b>	<b>291</b>
C.1	Machine Translation Post-editing for Typologically Diverse Languages	291
C.1.1	Subject Information . . . . .	291
C.1.2	Translation Guidelines . . . . .	291
C.1.3	Details on Document Selection and Preprocessing . . . . .	293
C.1.4	Other Measurements . . . . .	295
C.1.5	Data Filtering and Feature Significance . . . . .	301
C.2	Word-level Quality Estimation for Machine Translation Post-editing .	303
C.2.1	Filtering Details for QE4PE Data . . . . .	303
C.2.2	Additional Details and Statistics . . . . .	303
C.3	Unsupervised MT Error Detection and Human Disagreement . . . . .	314
C.3.1	Full Results . . . . .	314
<b>Summary</b>		<b>321</b>
<b>Samenvatting</b>		<b>323</b>

# CHAPTER 1

---

## Introduction

In recent years, language models have undergone a significant transformation, going from simple research prototypes producing barely coherent text to becoming a cornerstone of modern technological infrastructure. This success stems in large part from the remarkable ability of large neural networks such as the transformer (Vaswani et al., 2017) to learn rich representations of language—and by extension, our world and society—from staggering amounts of text. Yet, the complex and deeply intertwined structure that renders these systems so powerful is also the main culprit behind their opacity. The inner workings of neural networks remain notoriously difficult to interpret, and the lack of transparency in their decision-making processes has raised serious concerns about their reliability and fairness in high-stakes applications (Rudin, 2019).

These circumstances have led to a growing interest in *interpretability*—a field closely aligned with the broader area of explainable artificial intelligence (XAI), which seeks to develop methods and tools to understand how neural networks work and provide insights into their decision-making processes (Doshi-Velez and Kim, 2017; Li et al., 2022). In natural language processing (NLP), interpretability research has made significant strides by uncovering how language models encode and process factual knowledge and linguistic information (Tenney et al., 2019; Belinkov, 2022; Meng et al., 2022), revealing their use of context during generation (Clark et al., 2019; Ferrando et al., 2022b) and identifying the learned mechanisms underlying their capabilities (Elhage et al., 2021; Saphra and Wiegreffe, 2024).

While interpretability insights have earned broad recognition and influence within the NLP research community (Mosbach et al., 2024), critics have often pointed out that these findings rarely translate into actionable improvements for real-world systems (Räuker et al., 2023; Rai et al., 2024; Hendrycks and Hiscott, 2025). Most interpretability work today focuses on identifying subnetworks and mechanisms responsible for specific tasks inside language models (Ferrando et al., 2024; Sharkey et al., 2025), yet few

studies have put interpretability insights in relation to end-users' needs and desires (Ehsan et al., 2021), despite their crucial role in determining the practical usefulness of interpretability findings (Ehsan et al., 2024). This disconnect stems from a fundamental divide between research communities: most AI interpretability researchers pursue theoretical understanding of complex systems, while human-computer interaction (HCI) researchers prioritize actionable insights and practical applications.

A prime example of this disconnect can be found in the field of machine translation (MT), a long-standing area of research within NLP. MT researchers pioneered the use of neural language models for sequence generation tasks (Sutskever et al., 2014; Bahdanau et al., 2015), and were among the first to analyze their inner workings (Belinkov et al., 2017; Voita et al., 2019c; Rogers et al., 2020). Yet, despite the significant progress in the performance of MT systems across hundreds of languages over the past decade, the field has been remarkably slow to bring interpretability insights to the users of these systems, especially in the case of professional translators who work with these systems on a daily basis. Users of "classic" translation tools such as Google Translate are, to this day, simply presented with translations, without the possibility to personalize their tone or properties, quantify the model uncertainty in its response, or identify potential errors or alternative formulations. At the other extreme, when large language models like GPT-4 (OpenAI, 2023) eagerly offer eloquent justifications alongside their translations, these explanations may sound plausible but often fail to reflect the model's actual processing and context usage, resulting in plausible yet unfaithful rationalizations (Turpin et al., 2023).

This dissertation aims to bridge the gap between method-centric interpretability research and outcome-centric real-world machine translation applications. We develop novel methods to understand and control language model generation, then study how to integrate these advances effectively into human translation workflows. Our research spans three interconnected macro-themes: (1) understanding how language models exploit contextual information during generation, (2) controlling model generation for personalized translation outputs, and (3) integrating interpretability insights into human translation workflows. Our methodological contributions, empirical evaluations, and user studies demonstrate how insights from interpretability research can lead to meaningful impact in the way machine translation systems are used in real-world translation workflows.

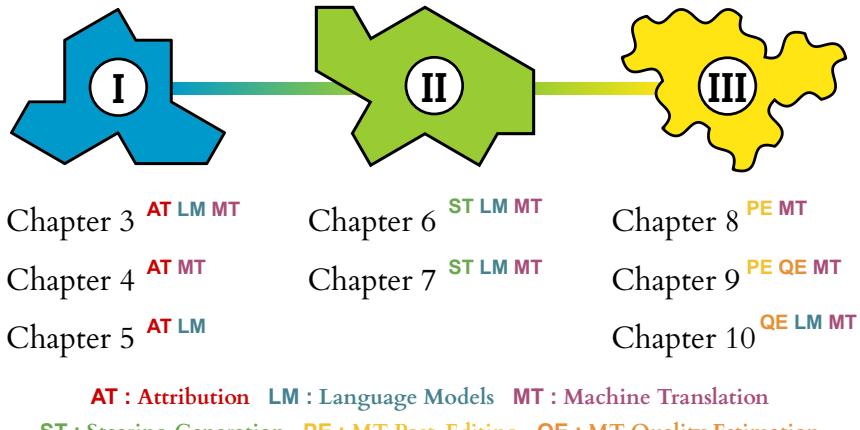


Figure 1.1 | Chapter guide for the three parts of this dissertation.

## 1.1 Outline and Contributions

The experimental chapters of this dissertation are organized into three parts, each addressing one of the research directions outlined above. Each part is composed of multiple chapters, each presenting a self-contained contribution or study related to the overarching theme. Figure 1.1 provides a visual overview of parts and chapters, highlighting for each chapter the topics introduced in detail in Chapter 2. Below, we summarize the contents, research questions and contributions of each part.

### Part I: Attributing Context Usage in Multilingual NLP

Part I establishes the foundational infrastructure and methodological frameworks for understanding how neural language models and machine translation systems process contextual information during generation. We begin with Inseq (Chapter 3), a toolkit that democratizes access to interpretability analyses of generative language models, providing the foundation for our investigations into context usage. Then, Chapter 4 introduces PECoRe, a data-driven framework for quantifying the plausibility of context usage in language models through the contrastive identification of context-sensitive tokens and contextual cues that influence their prediction. PECoRe is used to study context usage in context-aware machine translation systems, identifying failure cases stemming from an incorrect usage of context. Chapter 5 extends this analysis to modern large language models and retrieval-augmented generation settings with MIRAGE, adapting the PECoRe framework to demonstrate how model internals enable faithful

answer attribution in question answering. This part addresses two fundamental research questions:

**? Research Question 1 (RQ1)**

What are the conceptual and technical requirements for interpretability software tools enabling scalable and reproducible analyses into the inner workings of generative language models?

**? Research Question 2 (RQ2)**

How do language models and machine translation systems exploit contextual information during generation, and how can we quantify this usage in a faithful manner?

Part I's primary contributions include: (1) two open-source releases of the Inseq interpretability library; (2) the contrastive attribution tracing (CAT) method, a gradient-based alternative to causal intervention for efficiently identifying salient model components; (3) the PECoRe framework for context reliance attribution in language models, enabling data-driven exploration of context usage patterns in context-aware MT systems; and (4) an extended evaluation of context attribution for retrieval-augmented generation using MIRAGE, producing high quality citations of retrieved documents while ensuring greater faithfulness to the model's reasoning process.

## Part II: Conditioning Generation for Personalized Machine Translation

Part II moves from understanding context usage to actively controlling model generation for customized translation outputs. Across two chapters, we explore two paradigms to condition machine translation outputs—prompting-based methods and direct interventions in model processing—addressing the question:

**? Research Question 3 (RQ3)**

Are interpretability-based steering methods viable approaches for controllable machine translation? How do they compare with prompting-based methods in terms of their performance and their impact on models' internal mechanisms?

Chapter 6 pioneers the usage of prompting-based strategies for attribute-controlled

translation, while Chapter 7 connects generation conditioning to interpretability techniques, expanding the scope of our analysis from simple attributes in common domains to sophisticated personal styles in the challenging literary translation domain.

The core contributions of Part II include: (1) RAMP, a novel prompting methodology achieving strong performance in attribute-controlled translation across multiple languages and attributes without model fine-tuning; (2) the first comprehensive comparison of prompting versus interpretability-based steering for machine translation personalization; (3) a novel contrastive steering method using sparse autoencoder latents to achieve personalization accuracy comparable to prompting while preserving quality in literary translation; and (4) evidence that prompting and steering methods converge to similar mechanistic solutions, revealing fundamental principles of generation conditioning.

### Part III: Interpretability in Human Translation Workflows

Part III evaluates how interpretability insights can benefit human professionals who edit machine-translated content in a practical sense. We begin with DivEMT (Chapter 8), a study investigating the effectiveness of professional MT post-editing across a diverse set of mid-resourced languages, going beyond the one-size-fits-all analysis of high-resourced translation directions. This allows us to establish our human evaluation setup, providing valuable insights into the question:

? Research Question 4 (RQ4)

Does MT contribute positively to the productivity of professional translators across different languages? Which factors influence its effectiveness?

Building upon these insights, our second large-scale study QE4PE (Chapter 9) investigates how word-level error span highlights—including those derived from MT systems' uncertainty during generation—impact the productivity of professional translators and the quality of post-edited contents:

? Research Question 5 (RQ5)

How do word-level error highlights impact the productivity and editing choices of professional translators and the quality of resulting translations?

Chapter 10 concludes our human-centered investigation with a deeper analysis of

multiple uncertainty and interpretability-based word-level quality estimation methods. Such analysis allows us to assess how the performance of such techniques varies across different models, languages and human annotators:

**? Research Question 6 (RQ6)**

Can unsupervised error span detection methods reliably identify problems in machine-translated outputs? How does human label variation affect their performance, compared to traditional supervised approaches?

Part III contributions include (1) DivEMT, a cross-lingual post-editing dataset enabling controlled comparison of translator productivity across editing modalities and typologically diverse languages; (2) evidence that MT quality metrics fail to correlate with human post-editing productivity across languages, with productivity being heavily influenced by source-target language relatedness; (3) QE4PE, a comprehensive post-editing dataset containing error spans, behavioral editing metrics, and quality annotations from 42 professional post-editors for two translation directions; (4) evidence that error span highlights may reduce productivity but improve critical error detection; and (5) evidence that unsupervised quality estimation methods based on model internals can match state-of-the-art supervised approaches in both accuracy and downstream usability, revealing how subjective editing choice impact the evaluation of error span detection methods.

## 1.2 Scientific Output

This dissertation is the product of several research articles and open-source projects, which are categorized in the following sections.

### 1.2.1 Main Publications

The following articles represent the main contributions reflected in this thesis' experimental chapters, organized in their respective parts:<sup>1</sup>

#### Introduction and Background

---

<sup>1</sup>Shared first co-authorship is indicated by †.

- Ferrando, J., Sarti, G., Bisazza, A. and Costa-jussà, M. R. (2024). A Primer on the Inner Workings of Transformer-based Language Models. *Arxiv Preprint (Chapter 2)*

### Part I: Attributing Context Usage in Multilingual NLP

- Sarti, G., Feldhus, N., Sickert, L., van der Wal, O., Nissim, M. and Bisazza, A. (2023a). Inseq: An Interpretability Toolkit for Sequence Generation Models. In *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics (ACL Demo) (Chapter 3)*
- Sarti, G., Feldhus, N., Qi, J., Nissim, M. and Bisazza, A. (2024d). Democratizing Advanced Attribution Analyses of Generative Language Models with the Inseq Toolkit. In *Proc. of the 2nd World Conference on eXplainable Artificial Intelligence: Late-breaking works and demos (xAI Demo) (Chapter 3 and Chapter 4)*
- Sarti, G., Chrupała G., Nissim, M. and Bisazza, A. (2024c). Quantifying the Plausibility of Context Reliance in Neural Machine Translation. In *Proc. of the 12th International Conference on Learning Representations (ICLR) (Chapter 4)*
- Qi, J.<sup>†</sup>, Sarti, G.<sup>†</sup>, Fernández, R. and Bisazza, A. (2024). Model Internals-based Answer Attribution for Trustworthy Retrieval-Augmented Generation. In *Proc. of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Chapter 5)*

### Part II: Conditioning Generation for Personalized Machine Translation

- Sarti, G., Htut, P. M., Niu, X., Hsu, B., Currey, A., Dinu, G. and Nadejde, M. (2023b). RAMP: Retrieval and Attribute-Marking Enhanced Prompting for Attribute-Controlled Translation. In *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics (ACL) (Chapter 6)*
- Scalena, D.<sup>†</sup>, Sarti, G.<sup>†</sup>, Bisazza, A., Fersini, E. and Nissim, M. (2025). Steering Large Language Models for Machine Translation Personalization. *Arxiv Preprint (Chapter 7)*

### Part III: Interpretability in Human Translation Workflows

- **Sarti, G.**, Bisazza, A., Guerberof-Arenas, A. and Toral, A. (2022). DivEMT: Neural Machine Translation Post-Editing Effort Across Typologically Diverse Languages. In *Proc. of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (**Chapter 8**)
- **Sarti, G.**, Zouhar, V., Chrupała, G., Guerberof-Arenas, A., Nissim, M. and Bisazza, A. (2025b). QE4PE: Word-level Quality Estimation for Human Post-Editing. *Transactions of the Association for Computational Linguistics (TACL)* (**Chapter 9**)
- **Sarti, G.**, Zouhar, V., Nissim, M. and Bisazza, A. (2025a). Unsupervised Word-level Quality Estimation for Machine Translation Through the Lens of Annotators (Dis)agreement. In *Proc. of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (**Chapter 10**)

I led the conceptualization, implementation, experimental evaluation, and manuscript writing for each article for which I am the sole first author. For articles with shared first authorship, I co-led the conceptualization, experimental design, and manuscript writing. In Qi, Sarti, et al. (2024), I also implemented the API for experimental evaluation. The background in Chapter 2 adapts parts of our primer on transformer interpretability (Ferrando et al., 2024), for which I contributed by surveying the literature and writing content regarding transformer architecture, input attribution methods, steering approaches, and interpretability tools.

### 1.2.2 Open-source Contributions

Open-source software proved fundamental to this thesis, providing a solid foundation for conducting reproducible experimental work. Notably, all investigations we conducted employed solely open-source tools, models and datasets, despite the current popularity of proprietary language models. Each chapter provides links to all datasets, models, code, and demos to encourage scrutiny and foster further research.

My most notable contribution to the open-source research ecosystem is the **Inseq** toolkit, presented in Chapter 3, for which I serve as development lead. The library now counts 430+ Github stars and 80+ citations across international venues.

I also contributed to the development of the following open-source projects:

- The **Groningen Translation Environment (GroTE)**, a Gradio-based UI for machine translation post-editing supporting the live recording of behavioral

logs using the Hugging Face `datasets` hub and `spaces` ecosystem, developed with the help of Vilém Zouhar for the QE4PE study (Chapter 9). Available at <https://github.com/gsarti/grote> or via `pip install grote`.

- `gradio-highlightedtextbox`, a Svelte component for Gradio supporting text editing with highlighted spans, developed for collecting behavioral edit data in GROTE. Available at [https://huggingface.co/spaces/gsarti/gradio\\_highlightedtextbox](https://huggingface.co/spaces/gsarti/gradio_highlightedtextbox) or via `pip install gradio-highlightedtextbox`.
- `lab1`, a toolkit to facilitate token-level analyses of annotated texts with multiple edits and tokenization schemes, developed with the help of Vilém Zouhar for Chapter 10 analyses. Available at <https://github.com/gsarti/lab1> or via `pip install lab1`.
- **Interpreto**, a Python toolbox for concept-based interpretability analyses of language models maintained by the FOR/DEEL teams, which I helped design and develop as part of my visit to the IRT Saint Exupéry research institute in Toulouse, France. Interpreto is available at <https://github.com/FOR-sight-ai/interpreto> or via `pip install interpreto`.

The full set of open-source contributions, including demos, models, and datasets, are available on GitHub and 🐾 Hugging Face.

### 1.2.3 Other Research Contributions

Beyond this dissertation’s scope, my research output included projects organized around two main themes:

#### Advancing Italian natural language processing:

- Miaschi, A., Sarti, G., Brunato, D., Dell’Orletta, F. and Venturi, G. (2022). Prob-ing Linguistic Knowledge in Italian Neural Language Models across Language Varieties. *Italian Journal of Computational Linguistics (IJCoL)*
- Bianchi, F., Attanasio, G., Pisoni, R., Terragni, S., Sarti, G. and Balestri, D. (2023). Contrastive Language-Image Pre-training for the Italian Language. In *Proc. of the 9th Italian Conference on Computational Linguistics (CLiC-it)*
- Sarti, G. and Nissim, M. (2024). IT5: Text-to-text Pretraining for Italian Language Understanding and Generation. In *Proc. of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*

- **Sarti, G.**, Caselli, T., Nissim, M. and Bisazza, A. (2024b). Non Verbis, Sed Rebus: Large Language Models Are Weak Solvers of Italian Rebuses. In *Proc. of the 10th Italian Conference on Computational Linguistics (CLiC-it)*
- **Sarti, G.**, Caselli, T., Bisazza, A. and Nissim, M. (2024a). EurekaRebus – Verbalized Rebus Solving with LLMs: A CALAMITA Challenge. In *Proc. of the 10th Italian Conference on Computational Linguistics (CLiC-it)*
- Ciaccio, C., **Sarti, G.**, Miaschi, A. and Dell'Orletta, F. (2025). Crossword Space: Latent Manifold Learning for Italian Crosswords and Beyond. In *Proc. of the 11th Italian Conference on Computational Linguistics (CLiC-it)*

### Interpreting the inner workings of generative language models:

- Langedijk, A., Mohebbi, H., **Sarti, G.**, Zuidema, W. and Jumelet, J. (2024). DecoderLens: Layerwise Interpretation of Encoder-Decoder Transformers. In *Findings of the North American Chapter of the Association for Computational Linguistics (NAACL Findings)*
- Edman, L., **Sarti, G.**, Toral, A., van Noord, G. and Bisazza, A. (2024). Are Character-level Translations Worth the Wait? Comparing ByT5 and mT5 for Machine Translation. *Trans. of the Association for Computational Linguistics (TACL)*
- Scalena, D., **Sarti, G.** and Nissim, M. (2024). Multi-property Steering of Large Language Models with Dynamic Activation Composition. In *Proc. of the 7th Workshop on Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP)*
- Ghasemi Madani, M. R., Gema, A. P., **Sarti, G.**, Zhao, Y., Minervini, P. and Passerini, A. (2025). Noiser: Bounded Input Perturbations for Attributing Large Language Models. In *Proc. of the 2nd Conference on Language Modeling (CoLM)*
- Candussio, S., Saveri, G., **Sarti, G.** and Bortolussi, L. (2025). Bridging Logic and Learning: Decoding Temporal Logic Embeddings via Transformers. In *Proc. of the European Conference on Machine Learning and Principles of Knowledge Discovery in Databases (ECML-PKDD)*
- Islam, K. I. and **Sarti, G.** (2025). Reveal-Bangla: A Dataset for Cross-Lingual Multi-Step Reasoning Evaluation. In *Proc. of the 2nd Workshop on Bangla Language Processing (BLP)*.

Fin also had the privilege of co-organizing the BlackboxNLP 2025 workshop<sup>2</sup>—the leading venue for NLP interpretability work—and contributing to its shared task on benchmarking mechanistic interpretability methods for circuit localization and causal variable identification in large language models:

- Belinkov, Y., Mueller, A., Kim, N., Mohebbi, H., Chen, H., Arad, D., **Sarti, G.** (2025). Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP.
- Arad, D., Belinkov, Y., Chen, H., Kim, N., Mohebbi, H., Mueller, A., **Sarti, G.**, Tutek, M. (2025). Findings of the BlackboxNLP 2025 Shared Task: Localizing Circuits and Causal Variables in Language Models. In *Proc. of the 8th Workshop on Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP)*

---

<sup>2</sup><https://blackboxnlp.github.io/2025>



# CHAPTER 2

---

## Background

The initial sections of this chapter adapt contents from *A Primer on the Inner Workings of Transformer-based Language Models* (Ferrando et al., 2024).

*Distress not yourself if you cannot at first understand the deeper mysteries of Spaceland. By degrees they will dawn upon you.*

– Edwin A. Abbott, *Flatland: A Romance of Many Dimensions*, 1884

This chapter provides a succinct introduction to various topics discussed in the experimental chapters of this dissertation. Rather than a comprehensive review of relevant literature, it aims to provide key background about the research presented in this manuscript.

In particular, Section 2.1 and Section 2.4 discuss the basic functioning of neural networks-based language models and machine translation (MT) systems, and introduce the machine translation task representing the core focus of this work. Section 2.2 and Section 2.3 provide an introduction to methods for attributing inputs and conditioning generation in language models, corresponding to the topics discussed in Part I and Part II. Finally, Section 2.5 and Section 2.6 dive deeper in the translation domain, providing an overview of how MT models are employed in the translation industry by human post-editors, and discussing techniques for automatically evaluating machine translation quality. These notions provide a valuable background to Part III, which focuses on the impact of interpretability insights on human translation workflows.

Beyond this background section, each experimental chapter briefly summarizes relevant literature to contextualize the research questions and findings.

## 2.1 From Neural Networks to Neural Language Models

*Neural networks* are computational models which integrate principles from statistical learning theory (Vapnik, 1995), consisting of interconnected nodes (*neurons*) organized in layers, where each connection has an associated weight. Formally, a neural network is a function  $\mathbf{f} : \mathcal{X} \rightarrow \mathcal{Y}$  that maps inputs  $\mathbf{x} \in \mathcal{X}$  to outputs  $\mathbf{y} \in \mathcal{Y}$ , where  $\mathcal{X}$  and  $\mathcal{Y}$  are the input and output feature spaces, respectively. The function  $\mathbf{f}$  is parameterized by weights  $\theta \in \Theta$ , which are typically learned from data through the training process described in Section 2.1.1. Individual neurons are functions parametrized by *weights*  $\mathbf{w} \in \mathbb{R}^d$  and *biases*  $b \in \mathbb{R}$ , which are combined to produce an output  $\sigma(\mathbf{w}^T \mathbf{x} + b)$ , where  $\sigma$  is a non-linear *activation function*. Thanks to *non-linearities*, sequences of neurons can learn to represent complex relations from input vector  $\mathbf{x}$ .<sup>1</sup>

### 2.1.1 Supervised Learning for Neural Networks

In the *supervised learning* paradigm, given a training dataset  $\mathcal{D}$  containing paired instances:

$$\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\} \in (\mathcal{X} \times \mathcal{Y})^n \quad (2.1)$$

where  $\mathbf{x}_i$  is a vector of input features and  $y_i$  is the expected output, a neural network is trained to learn a functional mapping  $\mathbf{f}$  from inputs  $\mathbf{x}$  to labels  $\mathbf{y}$  by minimizing the average value of a *loss function*  $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathcal{R}$ , such that  $\ell(\mathbf{f}(\mathbf{x}), \mathbf{y})$  quantifies the gap between predicted outcomes  $\tilde{y}$  and ground truth  $y$  over examples in  $\mathcal{D}$ . The function  $\mathbf{f}$  is parameterized by weights  $\theta \in \Theta$ , which are optimized during training so as to minimize the loss function. Such optimization is typically performed using some variant of stochastic gradient descent (SGD), in which iterative steps  $1, \dots, t, \dots, T$  are taken to update  $\theta$  in the direction of the negative gradient of the loss function with respect to the weights:

$$\theta_{t+1} \leftarrow \theta_t - \eta \nabla_\theta \ell(\mathbf{f}(\mathbf{x}_j; \theta_t), \mathbf{y}_j) \quad (2.2)$$

where  $\eta$  is a chosen *learning rate*, and  $\mathbf{x}_j$  and  $\mathbf{y}_j$  are a subset of randomly sampled input-output pairs from the training set  $\mathcal{D}$ , typically referred to as *mini-batch*. This iterative refinement of model parameters is repeated until convergence, i.e. until the model performance on a left-out validation set does not improve significantly, allowing

---

<sup>1</sup>More details on neural networks can be found in Goodfellow et al. (2016).

for a robust convergence to a local minimum of the loss function, even for non-convex problems and high-dimensional parameter spaces.

We commonly refer to the inference process going from input  $\mathbf{x}$  to output  $\mathbf{y}$  as *forward pass*, and to the process of computing gradients and updating model parameters as *backward pass*.

### 2.1.2 Transformer Neural Networks

Transformers (Vaswani et al., 2017) are a class of neural networks that have become the de-facto standard for most natural language processing tasks, constituting the core neural network architecture employed throughout this thesis’ experiments. In essence, a transformer consists of a sequence of identical macro-layers, dubbed *blocks*, progressively contextualizing a sequence of input features  $\mathbf{Z} \in \mathbb{R}^{S \times d}$ , where  $S$  is the sequence length and  $d$  is the size of each feature vector. Figure 2.1 illustrates the structure of a single transformer module, constituting the core of *decoder-only language models* such as GPT-3 (Brown et al., 2020) presented later in Section 2.1.3. Notably, the transformer architecture is characterized by its ability to process input sequences in parallel, as opposed to recurrent models (Rumelhart and McClelland, 1987; Hochreiter and Schmidhuber, 1997), making it highly efficient for training on large datasets.

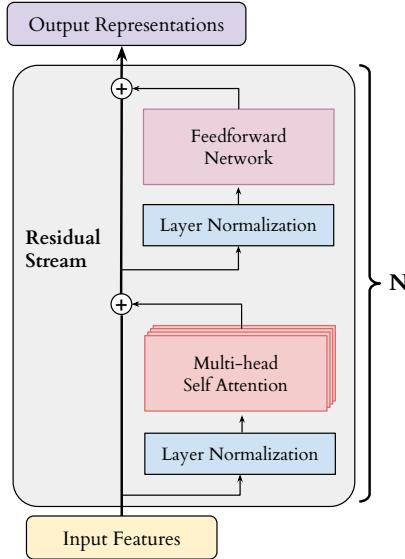
We now describe the main components of a transformer block in order of execution during the forward pass, using  $\mathbf{z}_i \in \mathbb{R}^d$  to denote the input representations at each component for sequence element  $i$ . This will be useful for explaining steering intervention and vocabulary projection methods used in Chapter 7 and Chapter 10, respectively.

**Layer normalization (LN).** The layer normalization operation, also known as LayerNorm (Ba et al., 2016), is a common approach for stabilizing the training process of deep neural networks. In practice, layer normalization applies the transformation:

$$\text{LN}(\mathbf{z}_i) = \frac{\mathbf{z}_i - \mu(\mathbf{z}_i)}{\sigma(\mathbf{z}_i)} \odot \gamma + \beta \quad (2.3)$$

where  $\mu, \sigma$  are the mean and the standard deviation of  $\mathbf{z}$ , and  $\gamma, \beta$  are learnable scale and bias parameters for the normalization. This operation helps to mitigate issues related to internal covariate shift, improving convergence during training. Recently, LayerNorm has been substituted with *RMSNorm* (Zhang and Sennrich, 2019), which removes the mean centering step and scales the input using the root mean square (RMS) statistic.

**Multi-head self-attention (MHSA).** The self-attention mechanism is the core component of the transformer architecture, allowing the model to contextualize its



**Figure 2.1** | An example transformer module with  $N$  blocks. We adopt the residual stream view of Elhage et al. (2021), with residual connections linearized to emphasize the read-write operations performed by attention and feed-forward network modules.

representations at each layer by combining information across the input sequence. While the original formulation of multi-head self-attention by Vaswani et al. (2017) involves a concatenation of attention head outputs before the final output projection, we follow here the more recent formulation by Kobayashi et al. (2021) and Elhage et al. (2021), which reformulates the attention output computation using the sum of individual attention heads, emphasizing the linear reading and writing operations within the attention computation.

Concretely, the self-attention module is composed by a series of  $H$  attention heads  $\text{Attn}_1, \dots, \text{Attn}_H$ , each computing the following weighted sum:

$$\text{Attn}_h(\mathbf{z}_i) = \sum_j \alpha_{ij}^h \mathbf{z}_j \mathbf{W}_V \mathbf{W}_O \quad (2.4)$$

Intuitively, the sharding of the attention mechanism into separate computations can be beneficial when processing the complex relations within different elements of the input sequence, for example, the lexical, syntactic and semantic dimensions of words in a text. The learnable weight matrices  $\mathbf{W}_V \in \mathbb{R}^{d \times d_h}$  and  $\mathbf{W}_O \in \mathbb{R}^{d_h \times d}$ , where  $d_h$  represents

the dimension of each head, can be combined into the so-called *output-value (OV) circuit* as  $\mathbf{W}_V \mathbf{W}_O = \mathbf{W}_{OV} \in \mathbb{R}^{d \times d}$ . For every key  $j$  given the current query position  $i < S$ , the corresponding *attention weight*  $\alpha_i^h$  is computed as:

$$\alpha_i^h = \text{softmax}\left(\frac{\mathbf{z}_i \mathbf{W}_Q (\mathbf{W}_K \mathbf{z})^T}{\sqrt{d_h}}\right) \quad (2.5)$$

Once again, the learnable weight matrices  $\mathbf{W}_Q \in \mathbb{R}^{d \times d_h}$  and  $\mathbf{W}_K \in \mathbb{R}^{d \times d_h}$  can be combined as the *query-key (QK) circuit*  $\mathbf{W}_Q \mathbf{W}_K^T = \mathbf{W}_{QK} \in \mathbb{R}^{d \times d}$ . This decomposition enables a view of QK and OV circuits as the units responsible for reading from (QK) and writing to (OV) the residual stream. Finally, the attention block output is the sum of individual attention heads:

$$\text{Attn}(\mathbf{z}_i) = \sum_{h=1}^H \text{Attn}_h(\mathbf{z}_i) \quad (2.6)$$

**Residual connection.** The introduction of residual connections (He et al., 2016) in the transformer architecture allows the model to learn identity mappings more easily, facilitating the training of deeper networks and avoiding the *vanishing gradient* problem (Hochreiter, 1998). A residual connection is commonly applied to the output of the self-attention module, resulting in:

$$\text{ResAttn}(\mathbf{z}_i) = \text{Attn}(\text{LN}(\mathbf{z}_i)) + \mathbf{z}_i \quad (2.7)$$

**Feedforward network (FFN).** The feedforward network (FFN) in the transformer block is composed of two learnable weight matrices<sup>2</sup>:  $\mathbf{W}_{\text{in}} \in \mathbb{R}^{d \times d_{\text{ffn}}}$  and  $\mathbf{W}_{\text{out}} \in \mathbb{R}^{d_{\text{ffn}} \times d}$ .  $\mathbf{W}_{\text{in}}$  reads from the residual stream state  $\mathbf{z}$ , and its result is passed through an element-wise non-linear activation function  $\sigma$ , producing a set of *neuron activations*. These get transformed by  $\mathbf{W}_{\text{out}}$  to produce the output  $\text{FFN}(\mathbf{z})$ :

$$\text{FFN}(\mathbf{z}_i) = \sigma(\mathbf{z}_i \mathbf{W}_{\text{in}}) \mathbf{W}_{\text{out}} \quad (2.8)$$

The FFN operation was compared to a retrieval step from a key-value memory (Geva et al., 2021), with keys stored in columns of  $\mathbf{W}_{\text{in}}$  acting as pattern detectors over the input sequence, and values in rows of  $\mathbf{W}_{\text{out}}$  being upweighted by respective neuron activation. The overall block structure from Figure 2.1 can then be summarized as:

---

<sup>2</sup>Bias terms can be omitted, following the practice of recent models such as Llama (Touvron et al., 2023)

$$\text{Block}(\mathbf{z}_i) = \text{FFN}\left(\text{LN}\left(\text{ResAttn}(\mathbf{z}_i)\right)\right) + \text{ResAttn}(\mathbf{z}_i) \quad (2.9)$$

We will henceforth use  $\mathbf{z}_i^l$  to denote the output of the  $l$ -th block for the  $i$ -th element of the input sequence for transformer models.

### 2.1.3 Transformer Language Models

A *language model* is a probabilistic model that can assign probabilities to sequences of tokens. Formally, given an input sequence  $\mathbf{X} = \langle t_1, \dots, t_S \rangle$  of  $S$  tokens, which in the case of natural language are typically words or subword units (Sennrich et al., 2016b) from a vocabulary  $\mathcal{V}$ , a language model  $f$  computes the probability of the sequence as the product of token-level conditional probabilities:

$$P(\mathbf{X}) = P(t_1, \dots, t_S) = \prod_{i=1}^S P(t_i | t_1, \dots, t_{i-1}) \quad (2.10)$$

Language models operating under such formulation are typically referred to as *auto-regressive* or *causal* language models (CLMs, or simply LMs), to differentiate them from *masked* language models (MLMs) trained to fill the blanks in a sequence (Devlin et al., 2019). While MLMs were the main object of analysis of early interpretability research on transformer models (Tenney et al., 2019; Clark et al., 2019; Rogers et al., 2020), this dissertation focuses solely on CLMs, which after the advent of ChatGPT<sup>3</sup> in 2022 became the dominant paradigm in the NLP and interpretability community. CLMs are typically decoder-only models, following the structure introduced in Section 2.1.2, or encoder-decoder models, such as the MT systems later discussed in Section 2.4.

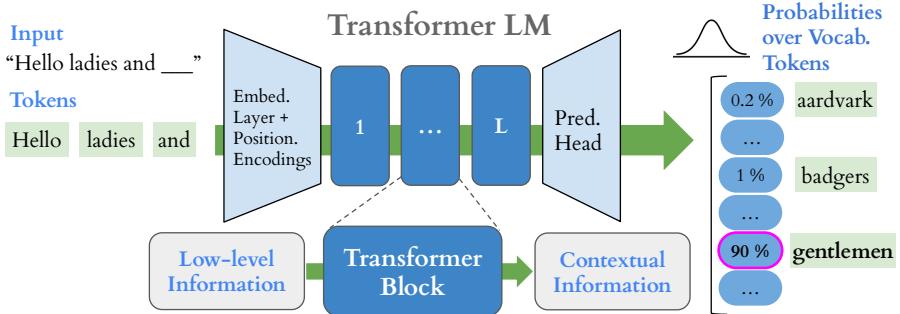
Importantly, LMs can be used for generating text by iteratively sampling from the probability distribution over the next token  $t_i$  given the previous tokens  $t_1, \dots, t_{i-1}$ , e.g. using the *greedy decoding* sampling method:

$$t_i^* = \arg \max_{t \in \mathcal{V}} P(\cdot | t_1, \dots, t_{i-1}) \quad (2.11)$$

This sampling process can be repeated autoregressively, i.e. by adding the selected token  $t_i^*$  to the input sequence, until a special *end-of-sequence* token is generated, or until a maximum sequence length is reached.

---

<sup>3</sup><https://openai.com/index/chatgpt>



**Figure 2.2 |** A transformer language model predicting the next word given a prefix.

We now turn to the additional components required to convert the generic transformer model presented in the previous section into a language model able to process and generate sequences of tokens. Figure 2.2 shows a stylized view of a transformer LM.

**Embedding layer.** The first component of a transformer language model is the *embedding layer*, which maps input tokens to continuous vector representations, known as *embeddings*. Word embeddings such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) revolutionized the field of natural language processing by exploiting *distributional semantics*, i.e. the fact that words which frequently appear in similar contexts should have similar meaning (Harris, 1954), to learn word representations end-to-end using gradient descent. In transformers, the *token embedding matrix*  $\mathbf{E} \in \mathbb{R}^{|\mathcal{V}| \times d}$ , where  $d$  is the size of the embedding vectors, and  $|\mathcal{V}|$  is the vocabulary size, is learned jointly with the rest of model parameters during training. The embedding layer maps each token  $t_i$  in the input sequence to its corresponding vector  $\mathbf{z}_i = \mathbf{E}[t_i]$ . The resulting sequence of embeddings  $\mathbf{Z} \in \mathbb{R}^{S \times d}$  corresponds to the input to the first transformer block. It is important to note that representations produced by  $\mathbf{E}$  are not contextualized, i.e. the same token  $t_i$  will always be mapped to the same vector  $\mathbf{z}_i$ , regardless of its meaning in the given sequence. For example, the word *ring* will always be mapped to the same vector, regardless of whether it is used as a noun or a verb. The transformer blocks are used to contextualize these representations, i.e. produce different vectors for the same token depending on the remainder of the sequence.

**Positional encodings.** While the sequential nature of language is an important factor in how we produce and process linguistic information, transformer models do not explicitly account for ordering across elements of the input sequence. For this reason, *positional encodings* injecting information about the position of each token in

the sequence are commonly used in transformer-based language models. The most basic positional encoding is a fixed sinusoidal encoding (Vaswani et al., 2017), which is added directly to the input embeddings. Recent models, however, employ rotary position embeddings, allowing for the encoding of both absolute and relative positions between tokens, and allowing the model to generalize to longer contexts beyond those seen during training (Su et al., 2024).

**Causal self-attention.** The self-attention mechanism in transformer language models is *causal*, meaning that the attention weights for each token  $t_i$  are computed only over the tokens preceding it in the sequence, i.e.  $t_1, \dots, t_{i-1}$ . This ensures that the model can only attend to past tokens when predicting the next token, preserving the auto-regressive nature of the model. The causal self-attention mechanism is implemented by masking out future tokens in the attention computation, ensuring that  $\alpha^h$  is computed only for  $j \leq i$  in Equation 2.4, and that only representations  $Z_{\leq i}$  are used to compute the key vector in Equation 2.5.

**Prediction head.** The prediction head of a transformer language models consists of a so-called *unembedding* matrix  $\mathbf{W}_U \in \mathbb{R}^{d \times |\mathcal{V}|}$  mirroring the initial embedding operation, sometimes accompanied by a bias. The last residual stream state  $\mathbf{z}_S^L$ , where  $L$  is the number of transformer blocks and  $S$  is the sequence length, gets transformed by this linear map converting the representation into a next-token distribution of logits, which is turned into a probability distribution via the softmax function:

$$P(\cdot | t_1, \dots, t_{i-1}) = \text{softmax}(\mathbf{z}_i^L \mathbf{W}_U) \quad (2.12)$$

In light of the residual stream view presented in Section 2.1.2, showing that different model components read from and write to the residual stream, it is natural to believe that the predictions derived by applying the unembedding matrix to the final residual stream state  $\mathbf{z}_S^L$  are the product of an *iterative refinement* across model components (Jastrzebski et al., 2018). The *logit lens* method (nostalgebraist, 2020), which we study for error detection in Chapter 10, exploits this intuition to analyze how the model refines the prediction throughout the forward pass, by projecting intermediate residual stream states  $\mathbf{z}_S^l$ , with  $l < L$ , to the vocabulary space using  $\mathbf{W}_U$ .

**Language model pre-training.** Modern language models such as those employed in this thesis are typically *pre-trained* on large web corpora spanning billions or trillions of tokens using the next-token prediction objective, i.e. minimizing the cross-entropy loss between the next-token distribution predicted by the model and the next observed

token. This frames the language model training problem as an instance of supervised learning, which we presented in Section 2.1.1. Formally, given a minibatch  $D_t$  of corpus  $\mathcal{D}$  composed by sequences of tokens  $\mathbf{X}_k = \langle t_1, \dots, t_{S_k} \rangle$ , the loss for a single training step is computed as:

$$\mathcal{L}_{\text{step}} = -\frac{1}{|D_t|} \sum_{\mathbf{X}_k \in D_t} \sum_{i=1}^{S_k} \log P(t_i | t_1, \dots, t_{i-1}) \quad (2.13)$$

Concretely, this corresponds to maximizing the likelihood of the observed tokens given the context provided by the preceding tokens, while minimizing the likelihood of all other incorrect tokens.

**Language model post-training.** After pre-training, language models can be used for generating text given some context, but mostly lack the ability to perform specific tasks without being provided explicit examples, or respond to queries as conversational agents. For this reasons, all language models used for our experiments underwent additional *supervised fine-tuning* (SFT, also known as instruction tuning), allowing them to learn input-output mappings for realistic user queries beyond natural text occurrences in the pre-training corpus (Howard and Ruder, 2018; Sanh et al., 2022). The fine-tuning process still involves the same  $\mathcal{L}_{\text{step}}$  loss function over a smaller, curated set of demonstrations. Some of the models we study—such as the Gemma 2 models from Chapter 7 or the Zephyr model from Chapter 5—underwent an additional *reinforcement learning from human feedback* (RLHF) step, in which the model is fine-tuned to maximize the likelihood of human preferences over pairs of model generations, using a reward model trained on human preferences. This process is typically performed using Proximal Policy Optimization [PPO; Schulman et al. (2017)] or similar reinforcement learning algorithms. Unless otherwise specified, we use the term *language model* to refer to transformer language models that were first pre-trained and then fine-tuned, representing the main focus of this thesis.

## 2.2 Explaining Predictions with Input Attribution

Contrary to linear models, where learned coefficients directly correspond to the influence of their respective features towards predictions, neural networks’ outcomes cannot be directly interpreted due to the presence of multiple nonlinearities across layers, rendering the attribution of model prediction to individual input features non-trivial. *Input attribution* methods, also known as *feature attribution*, were introduced to address

this issue by providing a principled way to assign importance scores to input features, clarifying the rationales behind model decisions (Zeiler et al., 2011).

Formally, for a model  $\mathbf{f} \in \mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y}$ , given an input  $\mathbf{x} \in \mathcal{X}$ , we can define the attribution method  $\gamma$  as a functional:

$$\gamma : \mathcal{X} \times \mathcal{F} \rightarrow \mathbb{R}^{|\mathcal{X}|}$$

so that  $\mathbf{a}_{\mathbf{f}(\mathbf{x})} = \gamma(\mathbf{x}, \mathbf{f})$  is a vector of attribution scores quantifying the influence of each element of  $\mathbf{x}$  on the model predictive distribution  $\mathbf{f}(\mathbf{x})$ , with higher scores representing greater importance (Fel, 2024). It is worth noting that attribution methods can rely on one or more specific outcomes  $y \in \mathcal{Y}$  from the predictive distribution  $\mathbf{f}(\mathbf{x})$ , such as perturbation-based approaches (Covert et al., 2021), or simply rely on the flow of information within the model to identify important input elements (Abnar and Zuidema, 2020). We call the former methods *target-dependent*, and we discuss them further in Chapter 4.

### 2.2.1 Attribution Method Categories

We now briefly summarize common families of input attribution methods, which are employed throughout the first part of this thesis. An in-depth overview of input attribution techniques for natural language processing can be found in Madsen et al. (2022b).

**Gradient-based attribution** For neural network models like transformer LMs, gradients are a natural source of input saliency which can be exploited for attribution purposes (Simonyan et al., 2014; Li et al., 2016). A simple gradient-based attribution corresponds to a first-order Taylor expansion of the model at a point  $\mathbf{x}$ , expressed as  $\nabla \mathbf{f}(\mathbf{x}) \cdot \mathbf{x} + \mathbf{b}$ . The resulting gradient  $\nabla_{\mathbf{x}}^c \mathbf{f}$  captures intuitively the *sensitivity* of the model prediction  $c$  to each element in the input. In the case of transformer LMs,  $\nabla_{\mathbf{x}}^{t^*} \mathbf{f} \in \mathbb{R}^{S \times d}$ , i.e. every dimension of the input embedding is associated with a attribution score, and the logit of the top predicted token  $t^*$  is used as differentiation target for gradient computation.<sup>4</sup> These scores are generally aggregated at a token level to obtain a more intuitive overview of the influence of individual tokens. This is commonly done by taking the  $L^p$  norm of the gradient vector:

---

<sup>4</sup>Probability scores are commonly used as differentiation targets, see discussion in Bastings et al. (2022).

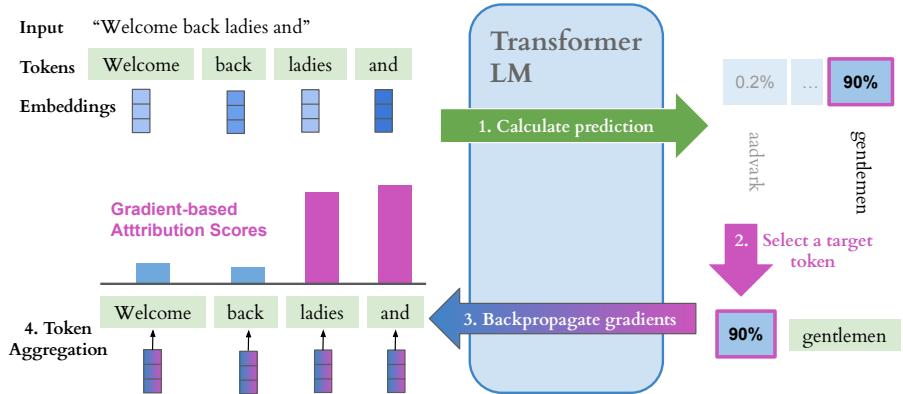


Figure 2.3 | Gradient-based attribution in a neural language model.

$$\text{Grad}_{\mathbf{f}(\mathbf{x}) \leftarrow t^*} = \|\nabla_{\mathbf{x}}^{t^*} \mathbf{f}\|_p \in \mathbb{R}^S \quad (2.14)$$

Figure 2.3 shows an example of gradient attribution on a language model. By taking the dot product between the gradient vector and the input embedding  $\nabla_{\mathbf{x}}^{t^*} \mathbf{f} \cdot \mathbf{x}$ , known as the *gradient × input* method, this sensitivity information can be converted to an importance estimate. More elaborate gradient-based attribution methods employ perturbations of the input embedding (Sundararajan et al., 2017; Smilkov et al., 2017) or ad-hoc gradient propagation rules (Bach et al., 2015; Achtibat et al., 2024) to filter noisy gradient information.

Gradient-based attribution methods are heavily used in the investigations of Chapter 3, Chapter 4 and Chapter 5, representing the majority of methods supported by the Inseq toolkit and the most effective approaches for contextual cues imputation in the PECoRe framework. Notably, gradient attribution can be exploited in a similar way to identify the importance of intermediate states  $\mathbf{z}$  in the model, as opposed to input representations  $\mathbf{x}$ , i.e. using  $\nabla_{\mathbf{z}}^{t^*} \mathbf{f}$ . The CAT method proposed in Chapter 3 case study adopts this attribution-based approach to locate factual knowledge across LM layers.

**Perturbation-based attribution** Another popular family of approaches estimates input importance by adding noise or ablating input elements and measuring the resulting impact on model predictions. For instance, the input token  $w_j$  at position  $j$  can be removed, and the resulting probability difference  $p(t^*|t_{<i}) - p(t^*_{\setminus w_j}|t_{<i})$ , where  $t^*$  is the predicted token for current sequence position  $i$  and  $j < i$ , can be used as an estimate for its importance. If the logit or probability given to  $w$  does not change,

we conclude that the  $i$ -th token has no influence. A multitude of perturbation-based attribution methods exist in the literature, such as those based on local surrogate models such as LIME (Ribeiro et al., 2016), or those derived from game theory like SHAP (Lundberg and Lee, 2017). Notably, some architecture-specific methods such as Value Zeroing (Mohebbi et al., 2023) have been proposed to mitigate the disruptive impact of perturbations on model behaviors. A comprehensive framework unifying various perturbation-based approaches is presented by Covert et al. (2021).

**Context mixing for attribution** Model internals such as the attention weights  $\alpha$  presented in Section 2.1.2 were initially proposed as possible explanations for model behavior (Bahdanau et al., 2015), but were found unfaithful in reflecting the actual predictive behavior of language models (Jain and Wallace, 2019; Bastings and Filipova, 2020). This is because, contrary to other approaches, they only accounted for the importance of specific model components, rather than a more general notion of saliency across the full model. However, recent methods have proposed more refined estimates of token contributions exploiting internals to quantify the information flow within LMs. Some of these alternatives include the use of the norm of value-weighted vectors and output-value-weighted vectors (Kobayashi et al., 2020, 2021), or the use of vectors' distances to estimate token contributions (Ferrando et al., 2022b). These methods result in a set of attribution scores  $\mathbf{a}_{\mathbf{f}(\mathbf{x})} \in \mathbb{R}^{S \times L}$ , marking the contribution of position-specific representation across all layers  $1, \dots, L$  of the model. These per-layer attributions reflecting context mixing patterns are often aggregated using techniques such as *rollout* (Abnar and Zuidema, 2020), resulting in one score per input token participating in the attention operation. Such context mixing approaches have shown competitive faithfulness compared to best gradient and perturbation-based methods, despite employing only a single forward pass to estimate contributions.

**Contrastive input attribution** An important limitation of input attribution methods for interpreting language models is that attributed output tokens belong to a large vocabulary space, often having semantically equivalent tokens competing for probability mass in next-word prediction (Holtzman et al., 2021). In this context, attribution scores are likely to misrepresent several overlapping factors such as grammatical correctness and semantic appropriateness driving the model prediction. Recent work addresses this issue by proposing a contrastive formulation of such methods, producing counterfactual explanations for why the model predicts token  $t^*$  instead of an alternative token  $t^\sim$ . Yin and Neubig (2022) extend the vanilla gradient method of Equation 2.14 to the contrastive setting as:

$$\text{ContGrad}_{\mathbf{f}(\mathbf{x}) \leftarrow t^*, t^\sim} = \nabla_{\mathbf{x}}^{t^* - t^\sim} \mathbf{f} \quad (2.15)$$

We employ this formulation in the PECoRE framework in Chapter 4 and its extension of Chapter 5 to identify salient context cues for generated tokens that were highly influenced by context.

### 2.2.2 Evaluating and Using Attribution Methods

**Plausibility and Faithfulness** The evaluation of input attribution methods can be operationalized in terms of various desiderata. *Plausibility*, also referred to as “human-interpretability” (Lage et al., 2019), is a measure of “*how convincing the interpretation is to humans*” (Jacovi and Goldberg, 2020), i.e. how well the salient tokens identified by an attribution method are in agreement with those selected by human annotators. It is important to note that plausibility does not imply *faithfulness*, i.e. how accurately the rationale reflects the true reasoning process of the model (Wiegreffe and Pinter, 2019), since a good explanation of model behavior might not align with human intuition. Consider the following sentence from the BLiMP corpus (Warstadt et al., 2020).

**x** = A report about the Impressionists has/\*have won the competition.

For the sentence to be grammatically correct, the verb *to have* must be correctly inflected as *has* to agree with the preceding noun *report*. Hence, to evaluate the plausibility of a language model for this example, the model is provided with the prefix  $\mathbf{x}'$  =“A report about the Impressionists”. Then, attribution scores are computed for every input token towards the prediction of *has* as the next token. Finally, we verify whether these scores identify the token *report* as the most important to predict *has*. We note that the selection of the pair *report-has* in the canonical procedure described above is entirely based on grammatical correctness, and other potential pairs not matching these constraints are not considered (e.g. the usage of *report* to predict *writing* instead of *has* as a likely continuation). This common procedure might also cause reasonable behaviors to be labeled as implausible. For example, the indefinite article *A* might be identified as the most important token to predict *has* since it is forcibly followed by a singular noun and can co-occur with *has* more frequently than *report* in the model’s training data. These limitations in the standard hypothesis-driven approach to plausibility evaluation motivate our proposal for PECoRE as a data-driven alternative in Chapter 4.

**Limitations of input attribution methods** While input attribution methods are commonly used to debug failure cases and identify biases in models’ predictions (McCoy

et al., 2019), popular approaches were shown to be insensitive to variations in the model and data generating process (Adebayo et al., 2018; Sixt et al., 2020), to disagree with each others’ predictions (Atanasova et al., 2020; Crabbé and van der Schaar, 2023; Krishna et al., 2024) and to show limited capacity in detecting unseen spurious correlations (Adebayo et al., 2020, 2022). Importantly, popular methods were found provably unreliable at predicting counterfactual model behavior in realistic settings (Bilodeau et al., 2024). Apart from theoretical limitations, perturbation-based approaches also suffer from out-of-distribution predictions induced by unrealistic noised or ablated inputs, and from high computational cost of targeted ablations for granular input elements.

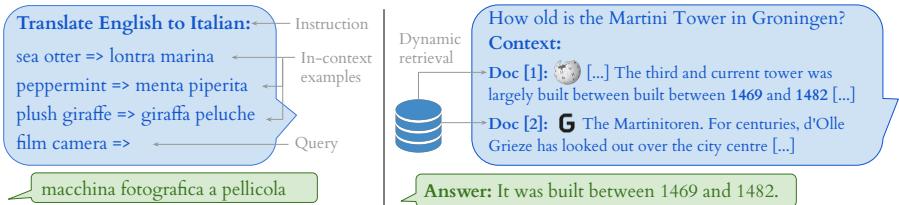
**Tools for input attribution** The `captum` library (Kokhlikyan et al., 2020) is part of the Pytorch ecosystem providing access to several gradient and perturbation-based input attribution methods for any Pytorch-based model, with the recent addition of utilities for simplifying attribution analyses of generative LMs (Miglani et al., 2023). Several `captum`-based tools provide convenient APIs for input attribution of transformer-based models, notably Transformers Interpret (Pierce, 2021), `ferret` (Attanasio et al., 2023) and `Ecco` (Alammar, 2021), which are mainly centered around language classification tasks. `SHAP` (Lundberg and Lee, 2017) is a popular toolkit mainly centered on perturbation-based input attribution methods and model-agnostic explanations for various data modalities. The `saliency` library<sup>5</sup> provides framework-agnostic implementations for mainly gradient-based input attribution methods, while `LIT` (Tenney et al., 2020) is a framework-agnostic tool providing a convenient set of utilities and an intuitive interface for interpretability studies spanning input attribution, concept-based explanations and counterfactual behavior evaluation. It notably includes a visual tool for debugging complex LLM prompts (Tenney et al., 2024). More recent low-level interpretability tools such as `nnsight` (Fiotto-Kaufman et al., 2025) also support attribution, without explicitly providing abstractions to facilitate its usage. `inseq`, which we introduce in Chapter 3 as part of this thesis’ contributions, is one of the most popular tools for input attribution of generative LMs, supporting advanced approaches for contrastive context attribution (Sarti et al., 2024c) and context mixing evaluation.

### 2.3 Conditioning Language Model Generations

This section describes the two main families of approaches for conditioning the behavior of language models during text generation. First, we present methods for modifying

---

<sup>5</sup><https://github.com/PAIR-code/saliency>



**Figure 2.4 | Left:** Few-shot prompting for English→Italian translation. **Right:** Retrieval-augmented generation for factual question answering. Relevant paragraphs are dynamically retrieved and infilled in the prompt using their similarity to the query to improve answer quality.

the input context by providing relevant information retrieved from external sources, or demonstrations of desired behavior, which we use in Chapter 5, Chapter 6, and 7. Then, we discuss approaches for modifying the model’s internal representations to achieve targeted interventions in the generation process, which we compare to prompting methods in Chapter 7.

### 2.3.1 Controlling Input Context

Large language models have become widely popular due to their ability to adjust their predictions in light of few examples or relevant information provided in an input context (*prompt*), without requiring additional training (Brown et al., 2020). Prompting LLMs to exploit their *in-context learning* skills has become pervasive in the NLP community, with much effort devoted to designing effective prompts for various tasks (Dong et al., 2024).

*Few-shot prompting* is an effective approach to adapt LLMs to new tasks by providing a few demonstrations of the desired behavior in the input context. For example, to perform a translation, a few source language examples can be provided in the prompt with their respective target language translations, and the model is expected to translate new source entries used as queries (Figure 2.4, left). *Zero-shot prompting* is a more challenging task, where the model is expected to perform well on a new task without any demonstrations, relying solely on its pre-trained knowledge. While effective, several studies highlighted the brittleness of prompting to unexpected factors such as the order of provided examples (Lu et al., 2022). In this thesis, we use few-shot prompting in our attribute-controlled translation experiments of Chapter 6 and our literary translation experiments of Chapter 7.

*Retrieval-augmented generation* (RAG) is a different approach for conditioning generation

where the model is provided with relevant context paragraphs retrieved on-the-fly from an external dataset, such as Wikipedia or a domain-specific corpus. This context is then used to inform the model’s predictions, allowing it to generate more accurate and relevant responses without relying solely on its potentially faulty pre-training knowledge (Figure 2.4, right). RAG has been shown to be effective in improving the factual accuracy of model outputs and reducing hallucinations (Lewis et al., 2020; Petroni et al., 2020). However, it is not directly obvious which retrieved paragraphs are motivating the model’s predictions, a challenge we address via input attribution in Chapter 5. Chapter 6 also employs a similarity retrieval component to control the examples selected for few-shot prompting, showing that example selection leads to better performances in machine translation with LLMs.

### 2.3.2 Controlling Model Representations

Techniques for conditioning model behavior by modifying the model’s internal representations are commonly referred to as *steering* methods, and often exploit the linear structure of model activations to achieve simple targeted interventions. Indeed, the *linear representation hypothesis* states that latent properties of interest—for example, the tone of a response—are encoded as linear subspaces of the representation space in language model activation (Park et al., 2023). Such property was already observed in early work on word embeddings (Mikolov et al., 2013), where the direction of the vector between two words was shown to encode their semantic relationship, e.g.  $\mathbf{z}_{\text{king}} - \mathbf{z}_{\text{man}} + \mathbf{z}_{\text{woman}} \approx \mathbf{z}_{\text{queen}}$ . Recent work highlighted the effectiveness of linear interventions on language models representations using directions identified by a *probing classifier*, i.e. a model  $\mathbf{p} : \mathbb{R}^d \rightarrow \mathcal{C}$  trained to predict a specific property of interest  $c \in \mathcal{C}$  from the intermediate representation of a trained transformer LM (Köhn, 2015; Gupta et al., 2015; see Belinkov, 2022, for a review). For instance, adding negative multiples of the sentiment direction ( $\mathbf{c}_{\text{sent}}$ ) to the residual stream, i.e. modifying the activation  $\mathbf{z}^l$  as  $\tilde{\mathbf{z}}^l \leftarrow \mathbf{z}^l - \alpha \mathbf{c}_{\text{sent}}$ , where here  $\alpha$  is a pre-selected *steering coefficient* controlling the intensity of the intervention, is sufficient to generate a text exhibiting the opposite sentiment label (Tigges et al., 2024). This simple procedure, known as *activation addition*, has become popular for conditioning desired attributes in model generations, including multiple properties at once (Scalena et al., 2024). Some of its variants omit probing classifiers and employ other unsupervised methods for computing feature directions, such as K-Means clustering of representations for examples showing a desired property (Zou et al., 2024), or mean difference between representations for positive and negative sets of demonstrations (Marks and Tegmark, 2024; Ardit et al., 2024).

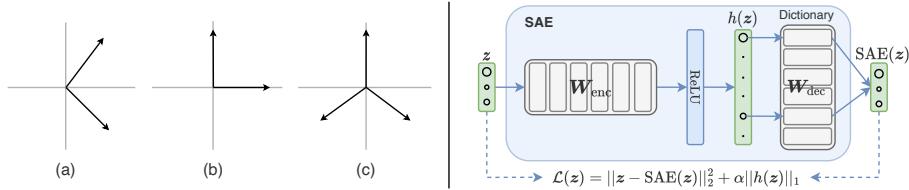
Wu et al. (2024) describe a broader framework for representation steering, proposing the use of *learnable interventions* for conditioning generation at specific steps with variable intensity. Formally, an intervention  $I$  can be defined as a tuple composed by an *intervention function*  $\xi : \mathbb{R}^d \rightarrow \mathbb{R}^d$  with learnable parameters, a set of input positions  $P \subseteq \{1, \dots, S\}$  that the intervention is applied to and the layer  $l$  at which the intervention is applied. This framework, dubbed *representation fine-tuning* (ReFT), allows to learn interventions overriding  $\mathbf{z}^l$  as:

$$z_i^l = \begin{cases} \xi(\mathbf{z}_i^l), & \text{if } i \in P \\ \mathbf{z}_i^l, & \text{otherwise} \end{cases} \quad (2.16)$$

The intervention function can be learned by minimizing the normal cross-entropy loss with a next token prediction objective, optimizing only the parameters of the intervention function. Activation addition (ActAdd) can then be described as a special case in this broader framework, where the intervention function  $\xi$  is constant and applied at all generation steps. In the experiments of Chapter 7, we use ActAdd and ReFT as baselines for our proposed steering method.

The final steering approach we discuss in this section involves the use of *sparse autoencoders*[SAEs; Huben et al. (2024)] for conditioning model behavior. SAEs have become widely adopted for analyzing the representations learned by transformer LMs thanks to their ability to address *polysemy*, i.e. the entanglement of multiple concepts within learned model representations. Indeed, neurons in transformer LMs were observed to activate on diverse and semantically distinct contexts, with concepts being encoded in a distributed manner across multiple units (Smolensky, 1986; Olah, 2023). In light of this, and given the disparity between the relatively low-dimensional representations learned by transformer LMs and the vast array of abilities they acquire during training, latent concept representations were speculated to be encoded in *superposition* across various model units (Arora et al., 2018), i.e. that multiple neurons jointly encode the presence of a single concept (Figure 2.5, left). A concrete example of this phenomenon is given by Elhage et al. (2022), where superposition is observed in presence of a long tail of sparse concepts in the training dataset.

A possible strategy to disentangle concepts in superposition involves finding an over-complete feature basis via dictionary learning (Olshausen and Field, 1997; Donoho and Elad, 2003). SAEs are simple autoencoder neural networks, i.e. models trained to reconstruct their input, that can be trained to reconstruct internal representations  $\mathbf{z} \in \mathbb{R}^d$  of a neural network exhibiting superposition. Their training objective encourages the model to learn a sparse coding of the input representation through an ad-hoc loss term,



**Figure 2.5 | Left:** Concepts encoded in a 2-dimensional parameter space. (a) Polysemytivity can be observed when concept do not align with the standard basis, i.e. they are encoded jointly by multiple units. (b) If concepts align perfectly with neurons, these neurons are *monosemantic*. (c) When the number of concepts exceeds the number of parameters, polysemytivity is inevitable and *superposition* is observed. **Right:** Sparse autoencoder (SAE) trained to reconstruct a model’s internal representations  $\mathbf{z}$ . Interpretive SAE concepts are found in rows of  $\mathbf{W}_{\text{dec}}$ . Biases are omitted for clarity.

resulting in a sparse dictionary of learned concepts. Huben et al. (2024) and Bricken et al. (2023) propose training SAEs on transformer LM representations using the form:

$$\begin{aligned} \text{SAE}(\mathbf{z}) &= h(\mathbf{z}) \mathbf{W}_{\text{dec}} + \mathbf{b}_{\text{dec}} \\ \text{with } h(\mathbf{z}) &= \sigma((\mathbf{z} - \mathbf{b}_{\text{dec}})\mathbf{W}_{\text{enc}} + \mathbf{b}_{\text{enc}}) \end{aligned} \quad (2.17)$$

using the loss function:

$$\mathcal{L}(\mathbf{z}) = \|\mathbf{z} - \text{SAE}(\mathbf{z})\|_2^2 + \alpha \|h(\mathbf{z})\|_1 \quad (2.18)$$

where  $\sigma$  is a non-linear activation function,  $\mathbf{W}_{\text{enc}}$  and  $\mathbf{W}_{\text{dec}}$  are the encoder and decoder learned weight matrices, respectively, and  $\alpha$  is a hyperparameter controlling the sparsity of the learned representation. The first term in Equation 2.18 is the *reconstruction term*, accounting for the quality of reconstruction, while the second term is the *sparsity term*, which promotes sparsity. The SAE architecture is illustrated in Figure 2.5 (right).

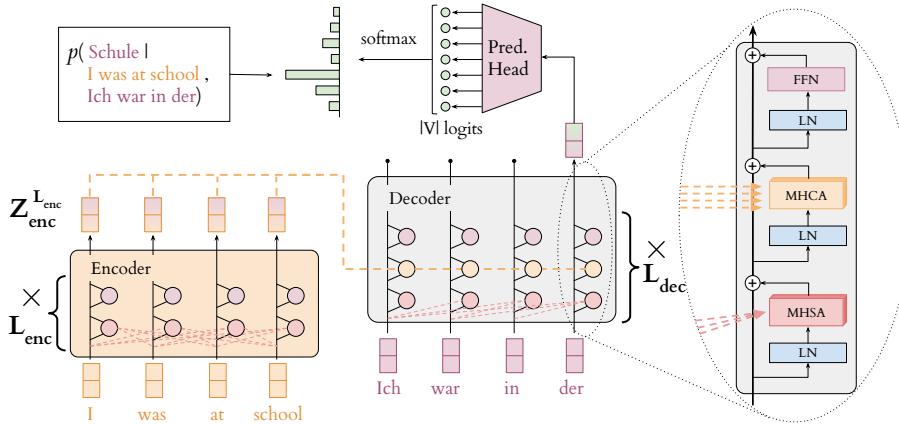
If  $h(\mathbf{z}) \in \mathbb{R}^m$  and  $m \gg d$ ,  $\mathbf{z}$  can be approximated as a sparse linear combination of the learned rows in the dictionary  $\mathbf{W}_{\text{dec}} \in \mathbb{R}^{m \times d}$ , ideally representing monosemantic concepts. Similarly to activation addition, these concepts can be used to steer model behavior by scaling them using a steering coefficient before reconstruction, resulting in a modified representation  $\tilde{\mathbf{z}}$ . We use a similar approach in our SAE-based steering method we present in Chapter 7.

## 2.4 Machine Translation

Machine translation is a long-standing task in natural language processing, with the goal of automatically translating text from a *source* language to another *target* language. In this section, we provide a brief overview of the evolution of machine translation approaches, describe how transformer LM architectures are commonly used for machine translation, and how such models can handle multiple languages and contextual information.

The history of machine translation can be summarized in three main phases. Between the 1960s and the 1980s, the first successes of machine translation were attained by **rule-based** systems exploiting various techniques, ranging from direct translation using dictionaries with a set of reordering rules to ambitious methods aiming to exploit an *interlingua* to act as a bridge when mapping meaning across languages (Hutchins, 2001). As for most rule-based methods, however, these approaches were limited by the need of ad-hoc rules, which could hardly account for less frequent and challenging settings. From the 1990s onwards, the **statistical** paradigm took foot by exploiting large bilingual corpora made available by the birth of the World Wide Web to train statistical language models parametrized as tables of co-occurrence probabilities (Och et al., 1999), with popular approaches aiming to segment challenging sentences into simpler phrases for ease of translation via co-occurrences (Koehn et al., 2003) or syntactic analysis (Hadiwinoto, 2017). In 2013, the advent of word embeddings coincided with the first MT systems based on continuous language representations parametrized by neural networks (Kalchbrenner and Blunsom, 2013), marking the advent of the **neural** MT (NMT) paradigm that remains the current state-of-the-art for machine translation. While the architecture of NMT systems has barely changed since the introduction of the transformer, as for most NLP tasks the introduction of large pre-trained language models has led to general-purpose models able to handle various translation-related task via light tuning and ad-hoc prompting (Alves et al., 2024).

Provided that machine translation involves the generation of a sequence of translated target tokens, it is straightforward to see how such task can fit well into the sequence-to-sequence framework adopted by neural language models. Given a sequence of tokens  $\mathbf{x} = (x_1, x_2, \dots, x_{S_s})$  in the source language  $s$ , a language model can be trained to generate a sequence of target tokens  $\mathbf{y} = (y_1, y_2, \dots, y_{S_t})$  in the target language  $t$  using the classic cross-entropy loss function. The transformer module we presented in Section 2.1.3 corresponds to the decoder-only architecture currently preferred for language modeling, involving a single stack of blocks. However, the original model proposed by Vaswani et al. (2017) followed the traditional *encoder-decoder* structure adopted in MT, with an additional dedicated component for encoding source information and



**Figure 2.6 |** Transformer encoder-decoder architecture for neural machine translation. The encoder processes the source sentence  $x$  and produces a sequence of contextualized representations, while the decoder generates the target sentence using causal self-attention (MHSA) and cross-attention (MHCA) mechanisms. The last decoder state is projected to the vocabulary space by the prediction head, and the next word is selected.

influencing the generation of target tokens.

The encoder-decoder transformer architecture for machine translation is illustrated in Figure 2.6. The encoder processes the source sentence  $x$  and produces a sequence of contextualized representations  $Z_{enc}^{L_{enc}} \in \mathbb{R}^{S_s \times d_{enc}}$  capturing the meaning of the source sentence. When generating the  $i$ -th token in the target sentence, every block of the decoder then attends to the target prefix  $y_{<i}$  using the self-attention module (MHSA) presented in Section 2.1.2, and complements this with a *multi-head cross-attention* (MHCA) mechanism integrating information from encoder representations  $Z_{enc}^{L_{enc}}$ . Functionally, the cross-attention module is identical to self-attention, but employs encoder representations to generate key and value vectors, while the query vectors are generated from the decoder representations.

While encoder-decoder transformers were traditionally trained from scratch on the machine translation task, the current state-of-the-art adapts pre-trained decoder-only LLMs with ad-hoc supervised tuning (Cui et al., 2025; Rei et al., 2024; Xu et al., 2024). Our experiments reflect this paradigm shift: initial MT experiments in Chapter 4, Chapter 8 and Chapter 9 employ traditional encoder-decoder, single-purpose translation models, while in Chapter 6 and Chapter 7 we generate translations by prompting general-purpose LLMs. Finally, Chapter 10 evaluates methods on both model types.

**Multilingual machine translation** Even before the advent of LLM-based translation systems, an important trend in MT research involved the training of massively multilingual MT (MMT) models capable of producing direct translations across hundreds of translation directions (Aharoni et al., 2019). Such approach was shown to bring improvements over previous methods requiring an intermediate translation step into a high-resource *pivot language* when two less-resourced languages were used as source and target (Kim et al., 2019a). MMT models are typically trained on large multilingual web corpora with similarity-matched sentence pairs in different languages (Schwenk et al., 2021), using special *language tags* such as `<eng_Latn>` as prefixes to mark source and target languages. After training, a translation into a specific language can be produced by prepending the respective language tag to the target sequence, biasing model generation towards tokens matching that language. This thesis makes ample use of encoder-decoder MMT models, such as mBART-50 (Tang et al., 2021), trained to translate from English to 50 languages (one-to-many MMT), M2M-100 (Fan et al., 2021), with many-to-many translation between 100 languages, and finally No Language Left Behind [NLLB; NLLB Team et al. (2024)], covering 200 languages in all directions. Decoder-only LLMs are generally trained on variable amounts of multilingual data<sup>6</sup>, and hence exhibit some degree of multilingual ability without additional MT tuning.

**Context-aware machine translation** Inter-sentential context is often fundamental for resolving discourse-level ambiguities during translation (Müller et al., 2018; Bawden et al., 2018; Voita et al., 2019b; Fernandes et al., 2023b). Traditional MT systems were trained at *segment level* due to their limited ability in handling long context, potentially losing important contextual information that spans beyond sentence boundaries, resulting in lower performances in realistic settings (Läubli et al., 2018; Toral et al., 2018a). *Context-aware MT* approaches aimed to address this limitation by incorporating document-level information to improve translation quality and consistency, leading to improved performance when translating cohesive discourse phenomena such as anaphora resolution, lexical cohesion, and maintaining consistent terminology within a document (Voita et al., 2018; Maruf and Haffari, 2018). Initial context-aware approaches for NMT employed methods ranging from concatenating multiple source sentences to employing hierarchical attention mechanisms that explicitly model document structure (Miculicich et al., 2018; Zhang et al., 2018). We use one such methods, namely concatenating context and current source text using a special `<brk>` tag, for the NMT

---

<sup>6</sup>Since the push towards proprietary model serving, details about the distribution of training data across languages in tech reports are often scarce.

models we analyze in Chapter 4. Recent LLM-based translation systems can naturally process longer contexts and maintain better consistency across document boundaries (Wang et al., 2023a; Briakou et al., 2024).

## 2.5 MT Post-Editing and Evaluation

The landscape of machine translation has undergone a fundamental transformation in recent decades, shifting from a tool primarily designed for professional translators to a technology accessed by millions of lay users worldwide (Savoldi et al., 2025). In this section, we review MT post-editing tools and practices, and discuss how MT outputs are evaluated by means of automatic metrics and human annotators.

### 2.5.1 Post-editing MT

Since the inception of MT technologies in professional translation workflow, human post-editing has been a crucial step to ensure quality and mitigate potential critical errors, especially for low-resource settings (Wagner, 1983; Church and Hovy, 1993). The industry distinguishes between two primary post-editing levels: *light post-editing*, which focuses on correcting only critical errors affecting comprehension while tolerating stylistic imperfections, and *full post-editing*, which aims to achieve human translation quality standards. The choice between these approaches involves trade-offs between effort investment and quality requirements, with light post-editing being faster while maintaining acceptable quality for many use cases (Plitt and Masselot, 2010). Seminal post-editing studies highlighted an increase in translators' productivity following MT adoption (Guerberof, 2009; Green et al., 2013; Läubli et al., 2013; Plitt and Masselot, 2010; Parra Escartín and Arcedillo, 2015). However, they also struggled to identify generalizable findings due to confounding factors like output quality, content domains, and high variance across language pairs and human subjects. With the advent of NMT, productivity gains of the new approach were extensively compared to those of statistical MT (Castilho et al., 2017; Bentivogli et al., 2016; Toral et al., 2018b; Läubli et al., 2019). Initial results were promising for NMT due to its better fluency and overall results. Moreover, translators were shown to prefer NMT over SMT for post-editing, although a pronounced productivity increase was not always present. In more recent times, various works explored the usage of adaptive MT systems that learn from post-editing feedback in real-time (Turchi et al., 2017; Karimova et al., 2018), with the goal of progressively reducing repetitive corrections and adapting to translator preferences. Notably, recent estimates confirm that human-machine collaboration can match or

even exceed the quality of human-only translations, with potential cost reductions estimated at around 60% the price of full human post-editing (Liu et al., 2024).

The main metric of evaluation for post-editing in the industry is *productivity*, often operationalized as the amount of source characters or word revised per minute. On the other hand, post-editing research often complements productivity measurements with *editing effort* alongside its *temporal*, *technical* and *cognitive* components (Krings, 2001), corresponding to editing time, number of keystrokes and pauses between keystrokes during the editing process, respectively. Importantly, the cognitive and temporal demands of post-editing were found to vary significantly depending on various factors, such as error types and user expertise. For example, Daems et al. (2017b) found that certain error categories have disproportionate impacts on post-editing effort, with adequacy errors often requiring more cognitive resources than fluency errors, even though the latter may be more immediately apparent to users (Martindale and Carpuat, 2018). Domain-specific considerations further complicate this landscape, as technical domains may tolerate certain stylistic variations while requiring precise terminology, whereas literary translation may prioritize creative renditions of meaning (Guerberof-Arenas and Toral, 2022).

Professional translators typically post-edit texts through *computer-assisted translation* (CAT) tools, which are interfaces designed to enhance human translators' productivity by providing access to keyboard shortcuts, quality estimation (which we discuss in Section 2.6) and other assistive technologies (Bowker, 2002). A common functionality of CATs is the integration of *translation memories* (TMs), which are bilingual databases storing previously translated content that can be retrieved and reused for similar segments, mimicking the functioning of early example-based MT systems (Garcia, 2009). Additional features often include terminology management systems (*termbases*) for maintaining consistency in technical terms and brand names, automatic text segmentation, and quality assurance modules such as spellcheckers for detecting errors and inconsistencies. Modern CAT tools have evolved from standalone desktop software to cloud-based platforms accessible via web browsers (Moran et al., 2014; Federico et al., 2014), with recent surveys indicating that 88% of professional translators use at least one CAT tool for their work.<sup>7</sup> While many CAT tools nowadays offer multiple advanced features, including LLM-based AI assistants, in our user studies of Chapter 8 and Chapter 9, we employ simple research-oriented interfaces with minimal text editing functionalities to ensure equal proficiency across subjects. In Chapter 8 we employ PET (Aziz et al., 2012), a simple desktop-based post-editing tool supporting various languages, while in

---

<sup>7</sup><https://go.proz.com/blog/cat-tool-use-by-translators-who-is-using>

Chapter 9 we use a custom-built web interface supporting editing over highlighted error spans.

### 2.5.2 MT Evaluation

The industrial context had historically an important influence on MT evaluation practices, encouraging researchers to focus on evaluation efficiency, combining automatic metrics with human assessment, and metrics that could provide concrete benefits when employed in professional translation workflows.

**Automatic MT Metrics.** Automatic evaluation metrics for machine translation have been widely adopted since the early 2000s, with the most popular metrics being BLEU (Papineni et al., 2002). BLEU is a simple and inexpensive metric measuring lexical similarity between a *candidate* translation  $\hat{y}$  and its given *reference*  $y$  as the number of  $n$ -grams  $G_n = \hat{y}_1, \dots, \hat{y}_n, \hat{y}_2, \dots, \hat{y}_{n+1}, \dots$  shared between them, normalized by the total n-gram count:

$$p_n(y, \hat{y}) = \frac{\sum_{s \in G_n} \min(C(s, \hat{y}), C(s, y))}{\sum_{s \in G_n} C(s, \hat{y})}$$

where  $C(s, y)$  is the count of n-gram  $s$  in sequence  $y$ . The complete BLEU score also incorporates a brevity penalty to discourage overly short translations. BLEU is computed at segment-level for an entire corpus of candidate and reference translations, and averaged to obtain a corpus-level score. Multiple variants of BLEU have been proposed to account for length bias, multiple references, with other metrics such as chrF (Popović, 2015) adopting similar lexicon-based approaches at the character level, or aligning n-grams across the two sequences (Banerjee and Lavie, 2005). Other lexical metrics such as the Translation Error Rate (Snover et al., 2006) or Word Error Rate (WER) have been used to connect the quality of the candidate sequence to the number of edits required to convert it into the reference, grounding the evaluation in post-editing technical effort. While these metrics provide rapid assessment of translation quality with minimal computational overhead, they suffer from several limitations: sensitivity to lexical variations that may not reflect translation quality differences, poor correlation with human judgments for high-quality neural MT outputs, and limited generalization across different writing systems (Bugliarello et al., 2020).

Following calls from the MT research community (Freitag et al., 2022), the limitations of lexical metrics led to the widespread adoption of *learned metrics* trained to predict translation quality from large amounts of annotated examples. Most of the widely

used learned MT metrics employ transformer-based encoder-only pretrained LMs such as BERT (Devlin et al., 2019) or the cross-lingual model XLM (Conneau and Lample, 2019). Among the most notable metrics, BLEURT (Sellam et al., 2020) is a BERT-based model using multi-task loss on synthetic data to perform regression of human quality judgments, while COMET (Rei et al., 2020) feeds source text, candidate and reference translation triples to a dual cross-lingual encoder structure that jointly learns to estimate quality and rank multiple candidate translations. In most of our MT evaluations we employ the COMET metric due to its excellent performance across hundreds of languages, which resulted in top-scoring submissions at multiple WMT metrics shared tasks (Rei et al., 2020, 2021, 2022a).<sup>8</sup> However, learned metrics introduce their own challenges, including non-trivial computational requirements, potential biases inherited from training data, and questions about generalization to out-of-domain content (Amrhein and Sennrich, 2022)

**Human evaluation of MT.** Human evaluation, despite its challenges due to inconsistencies across annotators, cultural and linguistic biases, and high costs, remains the gold standard for assessing machine translation quality, providing crucial insights that automatic metrics may fail to capture (Freitag et al., 2021a). Historically, human assessment of MT was centered around the notions of *adequacy* (also accuracy or fidelity), *comprehensibility* and *fluency* (or grammaticality) (White et al., 1994; Callison-Burch et al., 2007), with adequacy measuring how well the original meaning is conveyed, comprehensibility reflecting how understandable MT is without the original source, and fluency judging whether appropriate target grammar is employed (Popović, 2020). MT evaluation campaigns since 2017 adopted a *continuous direct assessment* (DA) of translation quality using scalar ratings—for example, using a 0-100 scale as in Graham et al. (2013)—or comparative ranking of multiple system outputs (Bojar et al., 2017). More recently, the introduction of the Multidimensional Quality Metric (MQM) (Lommel et al., 2013) has provided more structured evaluation protocols. MQM is an established framework allowing annotators to identify and categorize specific spans in a translated text as accuracy, fluency, and style issues, and assign them a level of severity (typically, a 3-way classification into *minor/major/critical*). Freitag et al. (2021a) experiments with various scoring configurations, resulting in the scoring formula:

$$\text{MQM} = (\# \text{ Major Err.} \times 5) + (\# \text{ Minor Err.} \times 1) + (\# \text{ Punct. Err.} \times 0.1)$$

---

<sup>8</sup>A comprehensive overview of MT metrics was released by Lee et al. (2023b).

with higher scores corresponding to worse translation, resulting in a high correlation with judgments from expert raters. However, such scheme has been criticized due to its potential length bias, with recent proposals for calibrated and non-linear scoring models accounting for similar issues (Lommel et al., 2024). An example description of MQM error categories and severity levels we employed for our study in Chapter 9 is presented in Table 9.1.

Recent evaluation campaigns such as WMT 2024 (Kocmi et al., 2024a) have increasingly adopted the MQM protocol for their evaluation, emphasizing in particular the importance of expert vs. non-expert annotators, with studies showing that translation professionals provide more consistent and reliable judgments compared to crowd-sourced annotations (Freitag et al., 2021a). The advent of large language models has introduced new challenges for human evaluation, as the quality gap between human and machine translation continues to narrow, requiring more fine-grained assessment criteria and larger annotator pools to achieve reliable results (Kocmi et al., 2024a). The main limiting factor towards the diffusion of the MQM evaluation protocol is its cost, since it involves a thorough annotation of error spans. Recently, the Error Span Annotation (ESA) protocol (Kocmi et al., 2024b) was introduced as a potential compromise between DA and MQM ratings, soliciting annotators to provide a 0-100 quality rating only after a light pass of error span identification, without requiring a full MQM error type categorization. The error annotation is intended to prime annotators to ground their quality judgments in empirical evidence, and ESA scores were observed to correlate strongly with MQM ones, while being 32% cheaper to obtain (Kocmi et al., 2024b). For this reason, we adopt a variant of the ESA protocol when conducting the quality assessment phase of our QE4PE study in Chapter 9. Zouhar et al. (2025) propose to use a language model to assist in the error span identification process, potentially further reducing the cost and effort involved in the ESA protocol.

## 2.6 Quality Estimation for MT

The automatic MT metrics presented in Section 2.5 require the use of a reference translation to measure the quality of a given candidate. While effective, these metrics cannot be employed to evaluate translation candidates on the fly, for example before presenting them to human post-editors, or as a ranking procedure in advanced decoding strategies (Rei et al., 2022b). Moreover, the presence of low-quality references can lead to biased evaluations of MT quality that do not reflect the translation quality without tying it to a specific gold standard (Freitag et al., 2023). **Quality estimation** metrics (QE), also known as *reference-free* MT metrics, are an alternative category of techniques

designed to address these limitations by predicting translation quality without requiring reference translations (Specia et al., 2018). Contrary to traditional MT evaluation, QE can be performed at various levels of granularity. On the one hand, when operating at the *segment* or *document* levels, QE methods typically return a score between 0 and 1 reflecting the overall quality of the translation, which can be then used to guide post-editors to focus on problematic segments (Tamchyna, 2021). On the other hand, *word-level* QE metrics can provide more granular information about translation issues, and typically operate by marking individual words with binary `OK`/`BAD` labels or, more recently, following the severity scheme introduced by the MQM framework.

Initial approaches to QE were mostly based on the uncertainty extracted from MT models (Blatz et al., 2004; Specia et al., 2009), but with time began focusing on supervised approaches involving ad-hoc model training (Turchi et al., 2013, 2014; Kepler et al., 2019; Thompson and Post, 2020, *inter alia*). Advances in segment- and word-level QE research are regularly assessed in annual WMT campaigns (Fomicheva et al., 2021; Zerva et al., 2022, 2024; Blain et al., 2023), where the best-performing QE systems have recently employed transformer-based language models trained to predict quality scores, in a fashion similar to reference-based metrics. In particular, reference-less counterparts to the COMET models were introduced for QE applications, including a smaller model for efficient inference (Rei et al., 2022b).

More recently, the widespread adoption of the MQM paradigm and the advances in LLM capabilities led to new QE metrics predicting quality at various granularity levels. Notably, Kocmi and Federmann (2023a) prompt GPT-4 with an annotation scheme mimicking MQM to produce fine-grained quality assessments, from which they derive a segment-level score, while Fernandes et al. (2023a) develop a similar AutoMQM framework using the PaLM-2 LLM. While these approaches usually employ proprietary models, Guerreiro et al. (2024) propose a state-of-the-art open-source QE model extending COMET to jointly predict quality estimation at the word and the sentence level, combining sentence-level and word-level error span prediction for improved explainability of results. XCOMET metrics come in a 3.5B (XL) and 10.7B (XXL) size and support both reference-based and reference-less usage, hence enabling usage for quality estimation purposes. Concretely, XCOMET models are transformer encoders fine-tuned from pre-trained XLMR encoders (Goyal et al., 2021) using a mix of sentence-level Direct Assessment scores and word-level MQM error spans. We use their resulting systems for our user study of Chapter 9 and our metric comparison in Chapter 10.

Aside from supervised models, a return to *unsupervised* methods exploiting models uncertainty and their internal mechanisms was brought on by Fomicheva et al. (2020).

In their work, such approaches were shown to rival state-of-the-art supervised QE models in predicting translation quality at the segment level. These methods typically rely on the model’s confidence in its predictions, often using metrics such as predictive probability or the entropy of the predictive distribution to mark low-confidence tokens as potential errors. The appeal of such methods lies in their efficiency, exploiting the knowledge of the MT model for error detection without requiring additional training on expensive human annotations. While such methods have been the object of multiple studies (Dale et al., 2023a; Xu et al., 2023b; Himmi et al., 2024; surveyed by Leiter et al., 2024), including a shared task dedicated to explainable QE metrics (Fomicheva et al., 2021), their evaluation was typically focused on segment-level evaluation quality, with word-level error spans being generally obtained by attributing the predictions of supervised segment-level metrics (Rubino et al., 2021; Rei et al., 2023). By contrast, recent work on LLMs evaluates various metrics to detect errors from the generator model, without additional systems involved, both at the sentence (Fadeeva et al., 2023) and at the token level (Fadeeva et al., 2024). Our evaluation of Chapter 10 involves various unsupervised metrics at the word level, employing the edits from our user studies of previous chapters as sources of word-level error spans to evaluate unsupervised word-level QE methods across multiple label sets. A notable technique for unsupervised QE is **Monte Carlo Dropout (MCD)** (Gal and Ghahramani, 2016). The dropout mechanism (Srivastava et al., 2014), commonly used for regularization during training, is employed at inference time by MCD to produce a set of noisy predictions from a unique model, approximating Bayesian inference. For a given input  $\mathbf{x}$ ,  $T$  forward passes are performed through the network. In each pass  $t \in T$ , a different random dropout mask  $\Theta_t$  is applied on model parameters, resulting in slightly different output probabilities  $p(\mathbf{x} | \Theta_t)$ . The set of  $T$  predictions  $\{p(\mathbf{x} | \Theta_1), \dots, p(\mathbf{x} | \Theta_T)\}$  can be seen as samples from an approximate posterior distribution. These can be used, for example, to quantify model uncertainty as the variance of the set of probabilities for a specific token. We employ such method, showing promising performances in our evaluation of Chapter 10, to produce unsupervised error highlights for our QE4PE user study in Chapter 9.

From a practical standpoint, QE methods are widely used in the translation industry for triaging automatic translations, with integrations in popular CAT tools to present users with segment-level quality scores (Tamchyna, 2021). While QE usage has been found helpful to increase the confidence and speed of human assessment (Mehandru et al., 2023; Zouhar et al., 2025), an incautious usage of these techniques can lead to a misplaced over-reliance on model predictions (Zouhar et al., 2021a). Moreover, the effectiveness of QE-assisted post-editing depends critically on the accuracy of quality

predictions, with inaccurate highlights potentially misleading translators and reducing overall productivity (Shenoy et al., 2021). Interfaces supporting word-level error highlights were developed for studying MT post-editing (Coppers et al., 2018; Herbig et al., 2020) and code reviewing (Sun et al., 2022; Vasconcelos et al., 2025), with results suggesting that striking the right balance of user-provided information is fundamental to improve the editing experience and prevent cognitive overload. Our user study of Chapter 9 is one of few works going beyond accuracy evaluations to measure the actual impact of word-level QE systems when integrated in human post-editing workflows.



## PART I

# Attributing Context Usage in Multilingual NLP



# CHAPTER 3

---

## Attributing Language Model Generations with the Inseq Toolkit

### Chapter Summary

This first experimental chapter presents the Inseq interpretability toolkit, which is employed for multiple analyses throughout this thesis. Inseq is a Python library that democratizes access to interpretability analyses of language models by enabling intuitive extraction of models' internal information and saliency scores throughout the generation process. After introducing Inseq design and features, we demonstrate its capabilities through applications that highlight gender biases in machine translation models and factual knowledge localization within the GPT-2 language model. Thanks to its extensible interface, which supports cutting-edge techniques, Inseq can drive future advances in explainable natural language generation, centralizing best practices and enabling reproducible model evaluations.

This chapter is adapted from the papers *Inseq: An Interpretability Toolkit for Sequence Generation Models* (Sarti et al., 2023a) and *Democratizing Advanced Attribution Analyses of Generative Language Models with the Inseq Toolkit* (Sarti et al., 2024d).

*As in manufacture so in science, retooling is an extravagance to be reserved for the occasion that demands it. The significance of crises is the indication they provide that an occasion for retooling has arrived.*

– Thomas S. Kuhn, *The Structure of Scientific Revolutions* (1970)

Recent years saw an increase in studies and tools aimed at improving our behavioral or mechanistic understanding of neural language models (Belinkov and Glass, 2019).

Many studies applied such techniques to modern deep learning architectures, including transformers (Vaswani et al., 2017), leveraging gradients (Baehrens et al., 2010; Sundararajan et al., 2017), attention patterns (Xu et al., 2015; Clark et al., 2019) and input perturbations (Zeiler and Fergus, 2014; Feng et al., 2018) to quantify input importance, often leading to controversial outcomes in terms of faithfulness, plausibility and overall usefulness of such explanations (Adebayo et al., 2018; Jain and Wallace, 2019; Jacovi and Goldberg, 2020; Zafar et al., 2021).

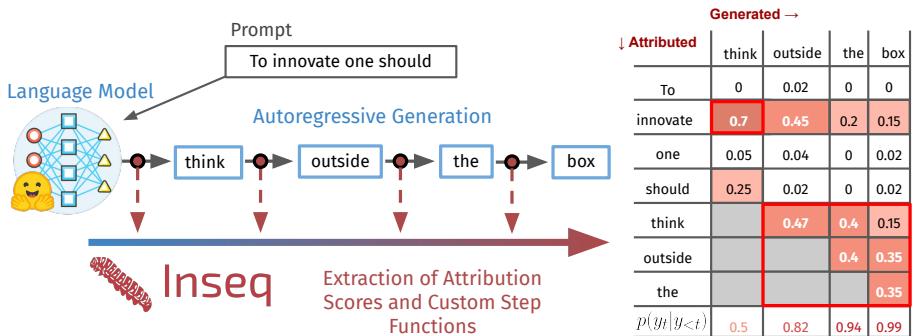
However, input attribution techniques have mainly been applied to classification settings (Atanasova et al., 2020; Wallace et al., 2020; Madsen et al., 2022a; Chrysostomou and Aletras, 2022), with relatively little interest in the more convoluted mechanisms underlying generation. Classification attribution is a single-step process resulting in one importance score per input token, often allowing for intuitive interpretations in relation to the predicted class. Sequential attribution<sup>1</sup> instead involves a computationally expensive multi-step iteration producing a matrix  $A_{ij}$  representing the importance of every input  $i$  in the prediction of every generation outcome  $j$  (Figure 3.1).

Moreover, since previous generation steps causally influence following predictions, they must be dynamically incorporated into the set of attributed inputs throughout the process. Lastly, while classification typically involves a limited set of classes and simple output selection (e.g., argmax after softmax), generation often operates with large vocabularies and non-trivial decoding strategies (Eikema and Aziz, 2020). These differences limited the use of input attribution methods for generation settings, with relatively few works improving attribution efficiency (Vafa et al., 2021; Ferrando et al., 2022a) and the informativeness of explanations (Yin and Neubig, 2022).

Having established a theoretical background on input attribution methods in Section 2.2, we introduce **Inseq**, a Python library that democratizes access to interpretability analyses of generative language models. Inseq centralizes access to a broad set of input attribution methods, sourced in part from the Captum (Kokhlikyan et al., 2020) framework, enabling a fair comparison of different techniques for all sequence-to-sequence and decoder-only models in the popular 🐄 `transformers` library (Wolf et al., 2020). Thanks to its intuitive interface, users can easily integrate interpretability analyses into sequence generation experiments with just 3 lines of code (Figure 3.2). Nevertheless, Inseq is also highly flexible, including cutting-edge attribution methods with built-in post-

---

<sup>1</sup>We use *sequence generation* to refer to all iterative tasks, including (but not limited to) natural language generation.



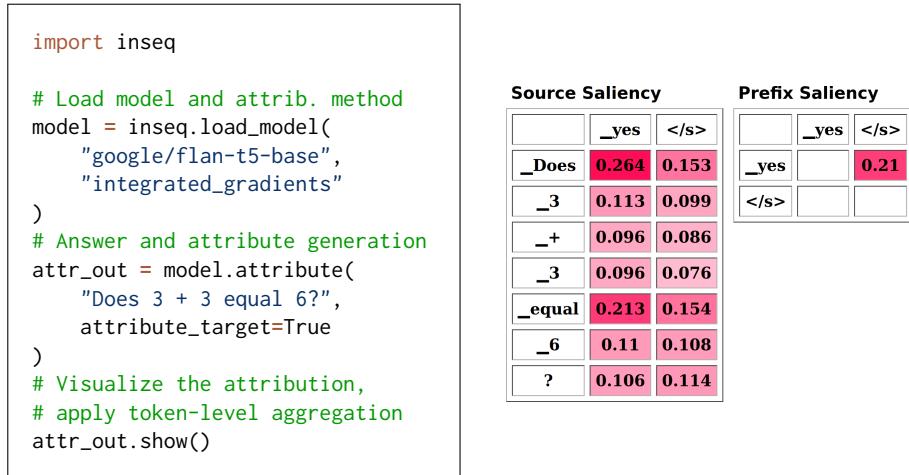
**Figure 3.1 |** Example of Inseq usage with a `transformers` causal language model. Given a prompt, attribution scores and next-step probabilities are extracted from the model at every generation step, with a final visualization aggregating values at the token level. Output attribution scores indicate that the model relies on the keyword “innovate” to initiate the idiomatic expression “think outside the box” with relatively low confidence ( $p = 0.5$ ). However, importance shifts to previous tokens in the idiom and confidence progressively grows throughout the generation.

processing features (Section 3.2.2), supporting customizable attribution targets and enabling constrained decoding of arbitrary sequences (Section 3.2.3).

In terms of usability, Inseq greatly simplifies access to local and global explanations, offering built-in support for a command-line interface (CLI), optimized batching that enables dataset-wide attribution, and various methods for visualizing, serializing, and reloading attribution outcomes and generated sequences (Section 3.2.4). Ultimately, Inseq aims to make sequence models first-class citizens in interpretability research and drive future advances in interpretability for generative applications.

### 3.1 Related Work

**Tools for NLP Interpretability** Although many post-hoc interpretability libraries were released recently, only a few support sequential input attribution. Notably, LIT (Tenney et al., 2020), a structured framework for analyzing models across modalities, and Ecco (Alammar, 2021), a library specialized in interactive visualizations of model internals. LIT is an all-in-one, GUI-based tool for analyzing model behaviors across entire datasets. However, the library does not provide out-of-the-box support for `transformers` models, requiring the definition of custom wrappers to ensure compatibility. Moreover, it has a steep learning curve due to its advanced UI, which can be inconvenient when working with a small number of examples. All these factors



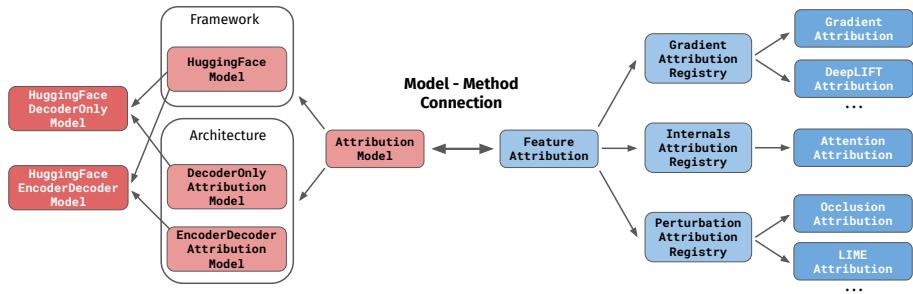
**Figure 3.2 |** Computing and visualizing attributions for Flan-T5 (Chung et al., 2024).

limit LIT usability for researchers working with custom models, needing access to extracted scores, or being less familiar with interpretability research. On the other hand, Ecco is closer to our work, being based on 🤗 `transformers` and having started to support encoder-decoder models concurrently with Inseq development. Despite a marginal overlap in their functionalities, the two libraries provide orthogonal benefits: Inseq’s flexible interface makes it especially suitable for methodical quantitative analyses involving repeated evaluations, while Ecco excels in qualitative analyses aimed at visualizing model internals. Other popular tools such as ERASER (DeYoung et al., 2020), Thermostat (Feldhus et al., 2021), `transformers-interpret` (Pierce, 2021) and ferret (Attanasio et al., 2023) do not support sequence models.

## 3.2 Design

Inseq combines sequence models sourced from 🤗 `transformers` (Wolf et al., 2020) and attribution methods mainly sourced from Captum (Kokhlikyan et al., 2020). While only text-based tasks are currently supported, the library’s modular design would enable the inclusion of other modeling frameworks, e.g. `fairseq` (Ott et al., 2019), and modalities (e.g. speech) without requiring substantial redesign. Optional dependencies include 🤗 `datasets` (Lhoest et al., 2021) and Rich.<sup>2</sup> Figure 3.3 presents the Inseq hierarchy of models and attribution methods. The model-method connection enables out-of-the-

<sup>2</sup><https://github.com/Textualize/rich>



**Figure 3.3 |** Inseq models and attribution methods. **Concrete** classes combine abstract **framework** and **architecture** attribution models classes, and are derived from abstract attribution methods' categories.

box attribution using the selected method. Framework-specific and architecture-specific classes enable the extension of Inseq to new modeling architectures and frameworks.

### 3.2.1 Guiding Principles

- **Research and Generation-oriented:** Inseq should support interpretability analyses of a broad set of sequence generation models without focusing narrowly on specific architectures or tasks. Moreover, the inclusion of new, cutting-edge methods should be prioritized to enable fair comparisons with well-established ones.
- **Scalable:** The library should provide an optimized interface to a wide range of use cases, models and setups, ranging from interactive attributions of individual examples using toy models to compiling statistics of large language models' predictions for entire datasets.
- **Beginner-friendly:** Inseq should provide built-in access to popular frameworks for sequence generation modeling and be fully usable by non-experts at a high level of abstraction, providing sensible defaults for supported attribution methods.
- **Extensible:** Inseq should support a high degree of customization for experienced users, with out-of-the-box support for user-defined solutions to enable future investigations into models' behaviors.

	<b>Method</b>	<b>Source</b>	$f(l)$
<b>G</b>	(Input $\times$ ) Gradient	Simonyan et al. (2014)	✓
	DeepLIFT	Shrikumar et al. (2016)	✓
	GradientSHAP	Lundberg and Lee (2017)	✗
	Integrated Gradients	Sundararajan et al. 2017	✓
	Discretized IG	Sanyal and Ren (2021)	✗
	<b>Sequential IG</b>	Enguehard (2023)	✗
<b>I</b>	Attention Weights	Bahdanau et al. (2015)	✓
<b>P</b>	Occlusion (Blank-out)	Zeiler and Fergus (2014)	✗
	LIME	Ribeiro et al. (2016)	✗
	<b>Value Zeroing</b>	Mohebbi et al. (2023)	✓
	<b>ReAgent</b>	Zhao and Shan (2024)	✗
<b>S</b>	(Log) Probability	-	
	Softmax Entropy	-	
	Target Cross-entropy	-	
	Perplexity	-	
	KL Divergence	-	
	Contrastive Logits/Prob. $\Delta$	Yin and Neubig (2022)	
	$\mu$ MC Dropout Prob.	Gal and Ghahramani (2016)	
	<b>PCXMI</b>	Fernandes et al. (2023)	
	<b>In-context PVI</b>	Lu et al. (2023)	
	<b>Top-p Tokens</b>		

**Table 3.1 |** Overview of gradient-based (**G**), internals-based (**I**) and perturbation-based (**P**) attribution methods and built-in step functions (**S**) available in Inseq.  $f(l)$  marks methods allowing for attribution of arbitrary intermediate layers. **Bolded methods** were introduced with Inseq v0.6.

### 3.2.2 Input Attribution and Post-processing

At its core, Inseq provides a simple interface for applying input attribution techniques to sequence generation tasks. We categorize methods in three groups, *gradient-based*, *internals-based* and *perturbation-based*, depending on their underlying approach to importance quantification.<sup>3</sup> Table 3.1 presents the complete list of supported methods. Aside from popular model-agnostic methods, Inseq notably provides built-in support for attention weight attribution and a range of cutting-edge methods not supported in any other toolkit, such as Discretized Integrated Gradients (Sanyal and Ren, 2021), Sequential Integrated Gradients (Enguehard, 2023), Value Zeroing (Mohebbi et al.,

<sup>3</sup>We distinguish between gradient- and internals-based methods to account for their difference in scores' granularity.

2023), and ReAGent (Zhao and Shan, 2024). Moreover, multiple methods support the importance attribution of custom intermediate model layers, simplifying studies on representational structures and information mixing in sequential models, as seen in our case study of Section 3.3.2.

**Source and target-side attribution** When using encoder-decoder architectures, users can set the `attribute_target` parameter to include or exclude the generated prefix in the attributed inputs. In most cases, this should be desirable to account for recently generated tokens when explaining model behaviors, such as when to terminate the generation (e.g. relying on the presence of `_yes` in the target prefix to predict `</s>` in Figure 3.2, right matrix). However, attributing the source side separately could be helpful, for example, to derive word alignments from importance scores.

**Post-processing of attribution outputs** Aggregation is a fundamental but often overlooked step in attribution-based analyses since most methods produce neuron-level or subword-level importance scores that would otherwise be difficult to interpret. Inseq includes several `Aggregator` classes to perform attribution aggregation across various dimensions. For example, the input word `Explanation` could be tokenized into two subword tokens `Expl` and `ation`, and each token would receive  $N$  importance scores, where  $N$  is the model embedding dimension. In this case, aggregators could first merge subword-level scores into word-level scores, and then merge granular embedding-level scores to obtain a single token-level score that is easier to interpret. Moreover, aggregation could prove especially helpful for long-form generation tasks such as summarization, where word-level importance scores could be aggregated to obtain a measure of sentence-level relevance. Notably, Inseq allows chaining multiple aggregators like in the example above using the `AggregatorPipeline` class, and provides a `PairAggregator` to aggregate different attribution maps, simplifying the conduction of contrastive analyses as in Section 3.3.1.<sup>4</sup>

### 3.2.3 Customizing generation and attribution

During attribution, Inseq first generates target tokens using `transformers` and then attributes them step-by-step. If a custom target string is specified alongside model inputs, the generation step is instead skipped, and the provided text is attributed by constraining the decoding of its tokens.<sup>5</sup> Constrained attribution can be used, among other things,

---

<sup>4</sup>See Section A.1.2 for an example.

<sup>5</sup>Users employing constrained decoding should be aware of its limitations in the presence of a high distributional discrepancy with natural model outputs (Vamvas and Sennrich, 2021b).

for contrastive comparisons of minimal pairs and to obtain model justifications for desired outputs.

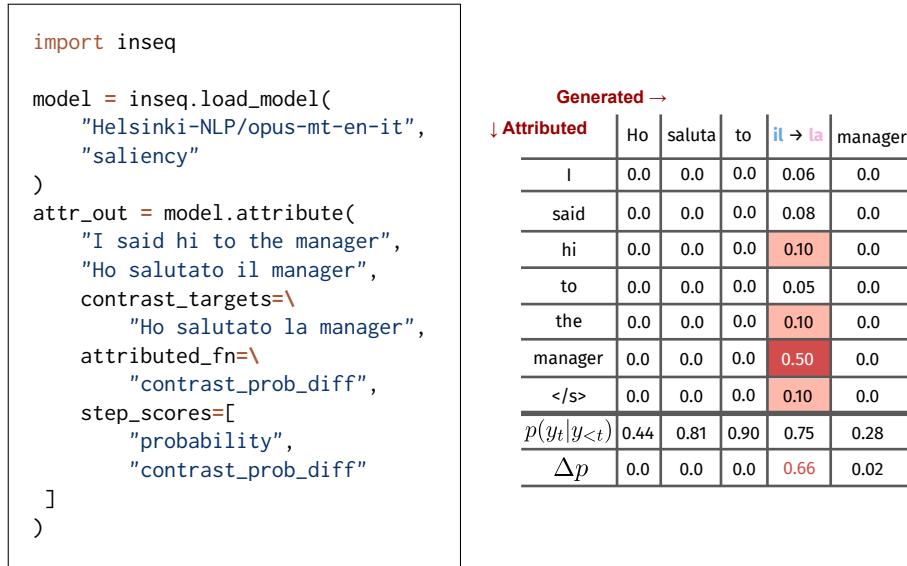
**Custom step functions** At every attribution step, Inseq can extract scores of interest (e.g. probabilities, entropy) that can be useful, among other things, to quantify model uncertainty (e.g. how likely the generated `_yes` token was given the context in Figure 3.2). We collectively refer to functions computing these scores as **step functions**. Inseq provides access to multiple built-in step functions (Table 3.1, S), enabling the computation of these scores, and allows users to create and register new custom ones. Step scores are computed together with the attribution, returned as separate sequences in the output, and visualized alongside importance scores (e.g. the  $p(y_t|y_{<t})$  row in Figure 3.1).

**Step functions as attribution targets** For methods relying on model outputs to predict input importance (gradient and perturbation-based), input attributions are commonly obtained from the model’s output logits or class probabilities (Bastings et al., 2022). However, recent work has shown the effectiveness of using targets, such as the probability difference of a contrastive output pair, to answer interesting questions like “What inputs drive the prediction of  $y$  rather than  $\hat{y}$ ? ” (Yin and Neubig, 2022). For example, the gradient  $\nabla(p(\text{barking}) - p(\text{crying}))$  given the prompt ““Can you stop the dog from \_\_\_” will highlight the role of the entity *dog* in selecting *barking*, disentangling the semantic component from grammatical correctness by providing a *crying* as grammatically valid choice. Figure 3.4 provides an example of such an approach for gender bias detection in machine translation. Inseq users can leverage any built-in or custom-defined step function as an attribution target, enabling advanced use cases like contrastive comparisons.

### 3.2.4 Usability Features

**Batched and span-focused attributions** The library provides built-in batching capabilities, enabling users to go beyond single sentences and attribute even entire datasets in a single function call. When the attribution of a specific span of interest is needed, Inseq also allows specifying a start and end position for the attribution process. This functionality greatly accelerates the attribution process for studies on localized phenomena (e.g. pronoun coreference in MT models).

**Alignment of contrastive options** Inseq supports customizable *word alignments*, i.e. indices aligning tokens in the original and contrastive generated texts, to support



**Figure 3.4** | Source-to-target attributions aggregated at the token level, indicating the importance of the stereotypical noun “manager” to generate the Italian masculine pronoun “il” (original) over the feminine “la” (contrastive case).

contrastive comparisons between texts of different lengths, including automatic alignments using the multilingual LaBSE encoder (Feng et al., 2022) to streamline their application.

**CLI, serialization and visualization** The Inseq library offers an API to attribute single examples or entire 📂 Datasets from the command line and save resulting outputs and visualizations to a file. Attribution outputs can be saved and loaded in JSON format, along with their respective metadata, to easily identify the provenance of the contents. Attributions can be visualized in the console or IPython notebooks and exported as HTML files.

**Quantized and distributed attribution** Supporting the attribution of large models is critical given recent scaling tendencies (Kaplan et al., 2020). All models that allow for quantization using `bitsandbytes` (Dettmers et al., 2022) can be loaded directly in 4-bit and 8-bit formats from 📂 `transformers`, and their attributions can be computed normally using Inseq at a fraction of the original computational cost.<sup>6</sup> Relatedly, Inseq

<sup>6</sup> `bitsandbytes 0.37.0` required for backward method, see Section A.1.3 for an example.

	Base		$\varphi \rightarrow \sigma$	
	$x_{\text{pron}}$	$x_{\text{occ}}$	$x_{\text{pron}}$	$x_{\text{occ}}$
$p(y_{\text{pron}})$	0.01		-0.44*	
$\nabla$	-0.16	0.25*	0.23*	-0.00
IG	-0.08	0.09	0.11	0.17
I $\times$ G	-0.11	0.22*	0.22*	-0.01

**Table 3.2 | Gender Bias in Turkish-to-English MT:** Kendall’s  $\tau$  correlation of MT model metrics with U.S. labor statistics. \* = Significant correlation ( $p < .05$ ).

is also compatible with the **Petals** framework (Borzunov et al., 2023), which supports gradient-based attribution across language models whose computation is distributed across multiple machines. This can alleviate the need for high-end GPUs to run LLMs, enabling the distributed computation of attribution scores.<sup>7</sup>

### 3.3 Case Studies

#### 3.3.1 Gender Bias in Machine Translation

In the first case study, we use Inseq to investigate gender bias in MT models. Studying the social biases embedded in these models is crucial to understanding and mitigating the representational and allocative harms they may engender (Blodgett et al., 2020). Savoldi et al. (2021) note that the study of bias in MT could benefit from explainability techniques to identify spurious cues exploited by the model and the interaction of different features that can lead to intersectional bias.

**Synthetic Setup: Turkish to English** The Turkish language uses the gender-neutral pronoun *o*, which can be translated into English as either *he*, *she*, or *it*, making it interesting to study gender bias in MT when associated with a language such as English, for which models will tend to choose a gendered pronoun form. Previous works have leveraged translations from gender-neutral languages to demonstrate the presence of gender bias in translation systems (Cho et al., 2019; Prates et al., 2020; Farkas and Németh, 2022). We repeat this simple setup using a Turkish-to-English MarianMT model (Tiedemann, 2020) and compute different metrics to quantify gender bias using Inseq.

We select 49 Turkish occupation terms verified by a native speaker (see Section A.1.1) and use them to infill the template sentence *O bir \_\_\_\_\_ (He/She is a(n) \_\_\_\_\_)*. For

<sup>7</sup>Tutorial: <https://inseq.org/en/latest/examples/petals.html>

each translation, we compute attribution scores for source Turkish pronoun ( $x_{\text{pron}}$ ) and occupation ( $x_{\text{occ}}$ ) tokens<sup>8</sup> when generating the target English pronoun ( $y_{\text{pron}}$ ) using Integrated Gradients (IG), Gradients ( $\nabla$ ), and Input  $\times$  Gradient ( $I \times G$ ).<sup>9</sup> We also collect target pronoun probabilities ( $p(y_{\text{pron}})$ ), rank the 49 occupation terms using these metrics, and finally compute Kendall’s  $\tau$  correlation with the percentage of women working in the respective fields, using U.S. labor statistics as in previous works (e.g., Caliskan et al., 2017; Rudinger et al., 2018). Table 3.2 presents our results.

In the **base case**, we correlate the different metrics with how much the gender distribution deviates from an equal distribution (50 – 50%) for each occupation (i.e., the gender bias irrespective of the direction). We observe a strong gender bias, with *she* being chosen only for 5 out of 49 translations and gender-neutral variants never being produced by the MT model. We find a low correlation between pronoun probability and the degree of gender stereotype associated with the occupation. Moreover, we note a weaker correlation for IG compared to the other two methods. For those, attribution scores for  $x_{\text{occ}}$  show significant correlations with labor statistics, supporting the intuition that the MT model will accord higher importance to source occupation terms associated to gender-stereotypical occupations when predicting the gendered target pronoun.

In the **gender-swap case** ( $\varphi \rightarrow \sigma$ ), we use the `PairAggregator` class to contrastively compare attribution scores and probabilities when translating the pronoun as *She* or *He*.<sup>10</sup> We correlate the resulting scores with the percentage of women working in the respective occupation and find strong correlations for  $p(y_{\text{pron}})$ , which supports the validity of contrastive approaches in uncovering gender bias.

**Qualitative Example: English to Dutch** We also qualitatively analyze biased MT outputs, showing how attributions can help develop hypotheses about models’ behavior. Table 3.3 (top) shows the  $I \times G$  attributions for English-to-Dutch translation using M2M-100 (Fan et al., 2021).

The model mistranslates the pronoun *her* into the masculine form *zijn* (his). We find that the wrongly translated pronoun exhibits high probability but does not associate substantial importance to the source occupation term *teacher*. Instead, we find good relative importance for the preceding word and *leraar* (male teacher). This suggests a strong prior bias for masculine variants, as shown by the pronoun *zijn* and the noun *leraar*, which may be a possible cause for this mistranslation. When considering

---

<sup>8</sup>For multi-token occupation terms, e.g., *bilim insanı* (scientist), the first token score was used.

<sup>9</sup>We set  $\Delta < 0.05$  for IG to ensure convergence. Token-level aggregation is performed using the L2 norm.

<sup>10</sup>An example is provided in Section A.1.2.

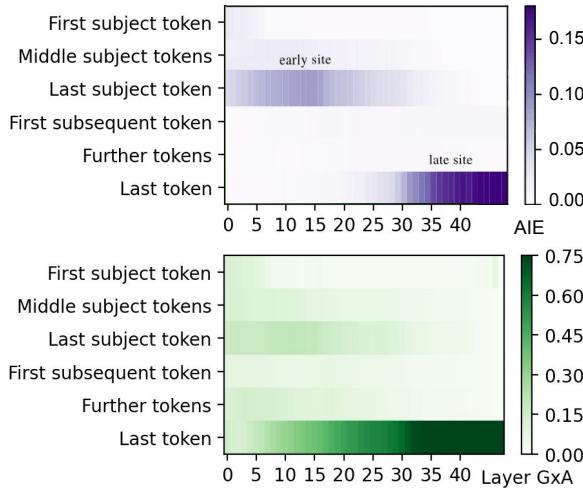
Source	De	leraar	verliest	zijn	baan
The	0.10	0.08	0.04	0.03	0.02
teacher	0.11	0.20	0.06	0.03	0.05
loses	0.11	0.09	0.25	0.07	0.07
her	0.15	0.09	0.10	0.21	0.07
job	0.10	0.08	0.08	0.10	0.24
Target	De	leraar	verliest	zijn	baan
De		0.23	0.05	0.06	0.04
leraar			0.17	0.13	0.03
verliest				0.18	0.08
zijn					0.26
$p(y_t)$	0.69	0.28	0.35	0.65	0.29
Source	De	$\sigma \rightarrow \emptyset$	verliest	haar	baan
The	0.00	-0.02	0.00	0.00	0.00
teacher	0.00	-0.05	-0.01	-0.01	-0.01
loses	0.00	-0.02	-0.01	-0.02	-0.01
her	0.00	-0.01	-0.01	-0.10	0.01
job	0.00	-0.02	-0.01	-0.02	-0.02
Target	De	$\sigma \rightarrow \emptyset$	verliest	haar	baan
De		-0.07	-0.01	0.01	-0.01
$\sigma \rightarrow \emptyset$			0.09	0.18	0.02
verliest				-0.03	0.00
haar					0.00
$\Delta p(y_t)$	0.00	-0.23	0.13	0.20	0.00

**Table 3.3 | Top:** Attribution of pronoun gender mistranslation using M2M-100. **Bottom:** Target attribution difference when swapping the target noun gender ( $\sigma \rightarrow \emptyset$ ) from *leraar* (male) to *leerkracht* (gender-neutral).

the contrastive example obtained by swapping *leraar* with its gender-neutral variant *leerkracht* (Table 3.3, bottom), we find increased importance of the target occupation in determining the correctly-gendered target pronoun *haar* (her). Our results highlight the tendency of MT models to attend inputs sequentially rather than relying on context, hinting at the known benefits of context-aware models for pronoun translation (Voita et al., 2018).

### 3.3.2 Locating Factual Knowledge inside GPT-2

For our second case study, we experiment with a novel attribution-based technique to locate factual knowledge encoded in the layers of GPT-2 1.5B (Radford et al., 2019). Specifically, we aim to reproduce the results of Meng et al. (2022), showing the influence of intermediate layers in mediating the recall of factual statements such as *The Eiffel Tower is located in the city of → Paris*. Meng et al. (2022) estimated the effect of network components in the prediction of factual statements as the difference in probability of a



**Figure 3.5 | Top:** Estimated causal importance of GPT-2 XL layers for predicting factual associations, as reported by Meng et al. (2022). **Bottom:** Average GPT-2 XL Gradient  $\times$  Layer Activation scores obtained with Inseq using contrastive factual pairs as attribution targets.

correct target (e.g. *Paris*), given a corrupted subject embedding (e.g. for *Eiffel Tower*), before and after restoring clean activations for some input tokens at different layers of the network. Apart from the obvious importance of final token states in terminal layers, their results highlight the presence of an early site associated with the last subject token playing an important role in recalling the network’s factual knowledge (Figure 3.5, top).

To verify such results, we propose a novel knowledge location method, which we name **Contrastive Attribution Tracing** (CAT), adopting the contrastive attribution paradigm of Yin and Neubig (2022) to locate relevant network components by attributing minimal pairs of correct and wrong factual targets (e.g. *Paris* vs. *Rome* for the example above). To perform contrastive attribution, we use the Layer Gradient  $\times$  Activation method, a layer-specific variant of Input  $\times$  Gradient, to propagate gradients up to intermediate network activations rather than reaching input tokens. The resulting attribution scores hence answer the question “*How important are layer L activations for prefix token t in predicting the correct factual target over a wrong one?*”. We compute attribution scores for 1000 statements taken from the Counterfact Statement dataset (Meng

et al., 2022) and present averaged results in Figure 3.5 (bottom).<sup>11</sup> Our results closely align with those of the original authors, providing additional evidence that attribution methods can be used to identify salient network components and guide model editing, as demonstrated by Dai et al. (2022).

We introduced the proposed CAT method shortly before the attribution patching technique by Nanda (2023). Together, these two methods represent the most efficient knowledge location techniques based on gradient propagation, with our approach requiring only a single forward and backward pass of the attributed model. Patching-based approaches, such as causal mediation (Meng et al., 2022), on the other hand, provide causal guarantees of feature importance at the price of being more computationally intensive. Despite lacking the causal guarantees of such methods, CAT can provide an approximation of feature importance and greatly simplify the study of knowledge encoded in large language model representations, thanks to its efficiency.

### 3.4 Conclusion

We introduced Inseq, a versatile and easy-to-use toolkit for interpreting sequence generation models. With many libraries focused on the study of classification models, Inseq is the first tool explicitly designed to analyze systems for tasks such as machine translation, code generation, and conversational applications. Researchers can easily add interpretability evaluations to their studies using our library to identify unwanted biases and interesting phenomena in their models’ predictions.

With the Inseq toolkit providing the foundational infrastructure for interpretability analysis, the following chapters will leverage the supported input attribution techniques to investigate context usage in context-aware machine translation systems Chapter 4 and multilingual language models for retrieval-augmented generation Chapter 5.

---

<sup>11</sup>Figure A.3 of Section A.1.3 presents some examples.

# CHAPTER 4

---

## Quantifying Context Usage in Neural Machine Translation

### Chapter Summary

This chapter investigates how context-aware machine translation models leverage contextual information. For this purpose, we introduce **Plausibility Evaluation of Context Reliance** (PECoRE), an end-to-end interpretability framework designed to quantify context usage in language models' generations. Our approach leverages model internals to contrastively identify context-sensitive target tokens in generated texts and link them to contextual cues justifying their prediction. We demonstrate the framework's effectiveness by assessing the plausibility of context-aware machine translation models, comparing model rationales with human annotations across several discourse-level phenomena. We integrate PECoRE in the Inseq toolkit API and apply it to unannotated model outputs to identify context-mediated predictions and highlight instances of (im)plausible context usage throughout generation.

This chapter is adapted from the paper *Quantifying the Plausibility of Context Reliance in Neural Machine Translation* (Sarti et al., 2024c). Section 4.6 is adapted from the case study in *Democratizing Advanced Attribution Analyses of Generative Language Models with the Inseq Toolkit* (Sarti et al., 2024d).

*An interpretation will be meaningful to the extent that it accurately reflects some isomorphism to the real world.*

– Douglas R. Hofstadter, Gödel, Escher, Bach: An Eternal Golden Braid (1979)

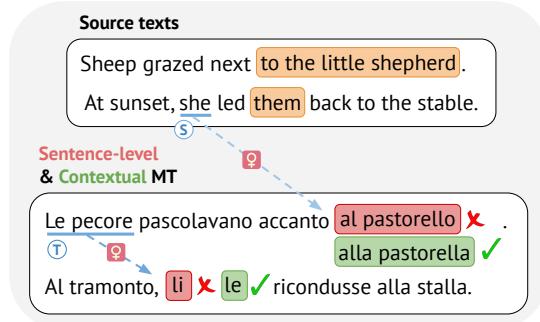
## 4.1 Introduction

Research in NLP interpretability defines various desiderata for rationales of model behaviors, i.e. the contributions of input tokens toward model predictions computed using input attribution (Madsen et al., 2022b). One such property is *plausibility*, corresponding to the alignment between model rationales and salient input words identified by human annotators (Jacovi and Goldberg, 2020). Low-plausibility rationales typically occur alongside generalization failures or biased predictions and can be helpful in identifying cases where models are “right for the wrong reasons” (McCoy et al., 2019).

However, while plausibility has an intuitive interpretation for classification tasks involving a single prediction, extending this methodology to generative language models presents several challenges. First, LMs have a large output space in which semantically equivalent tokens (e.g. “PC” and “computer”) are competing candidates for next-word prediction (Holtzman et al., 2021). Moreover, LMs’ generations are the product of optimization pressures to ensure independent properties such as semantic relatedness, topical coherence and grammatical correctness, which can hardly be captured by a single attribution score (Yin and Neubig, 2022). Finally, since autoregressive generation involves an iterative prediction process, model rationales could be extracted for every generated token. This raises the issue of *which generated tokens* can have plausible contextual explanations.

Recent attribution techniques for explaining language models incorporate contrastive alternatives to disentangle different aspects of model predictions (e.g. the choice of “meowing” over “screaming” for “*The cat is \_\_*” is motivated by semantic appropriateness, but not by grammaticality) (Ferrando et al., 2023; Sarti et al., 2023a). However, these studies circumvent the issues above by focusing their evaluation on a single generation step matching a phenomenon of interest. For example, given the sentence “*The pictures of the cat \_\_*”, a plausible rationale for the prediction of the word “are” should reflect the role of “pictures” in subject-verb agreement. While this approach can be helpful to validate model rationales, it confines plausibility assessment to a small set of handcrafted benchmarks where tokens with plausible explanations are known in advance. Moreover, it risks overlooking important patterns of context usage, including those that do not immediately match linguistic intuitions. In light of this, we suggest that identifying *which generated tokens* were most affected by contextual input information should be an integral part of plausibility evaluation for language generation tasks.

To achieve this goal, we propose a novel interpretability framework, which we dub **Plausibility Evaluation of Context Reliance (PECoRE)**. PECoRE enables the end-to-



**Figure 4.1** | Examples of sentence-level English→Italian translation with **lack-of-context errors** and their correct contextual counterpart. In the contextual case **context-sensitive source tokens** are disambiguated using source (S) or target-based (T) **contextual cues** to produce correct **context-sensitive target tokens**. PECoRE enables the end-to-end extraction of **cue-target pairs** (e.g. **she-allà pastorella**, **le pecore-le**).

end extraction of *cue-target token pairs* consisting of context-sensitive generated tokens and their respective influential contextual cues from language model generations, as shown in Figure 4.1. These pairs can uncover context dependence in naturally occurring generations and, for cases where human annotations are available, help quantify the plausibility of context usage in language models. Importantly, our approach is compatible with modern attribution methods using contrastive targets (Yin and Neubig, 2022), avoids relying on reference translations to avoid problematic distributional shifts (Vamvas and Sennrich, 2021b), and can be applied to unannotated inputs to identify context usage in model generations.

After formalizing our proposed approach in Section 4.3, we apply PECoRE to contextual machine translation to study the plausibility of context reliance in bilingual and multilingual MT models. While PECoRE can easily be used alongside encoder-decoder and decoder-only language models for interpreting context usage in any text generation task, we focus our evaluation on MT because of its constrained output space facilitating automatic assessment and the availability of MT datasets annotated with human rationales of context usage. We thoroughly test PECoRE on well-known discourse phenomena, benchmarking several context sensitivity metrics and attribution methods to identify cue-target pairs. We conclude by applying PECoRE to unannotated examples and showcasing some reasonable and questionable cases of context reliance in

MT model translations.<sup>1</sup>

In sum, we make the following contributions:

- We introduce PECoRE, an interpretability framework to detect and attribute context reliance in language models. PECoRE enables a quantitative evaluation of plausibility for language generation beyond the limited artificial settings explored in previous literature.
- We compare the effectiveness of context sensitivity metrics and input attribution methods for context-aware MT, showing the limitations of metrics currently in use.
- We apply PECoRE to naturally-occurring translations to identify interesting discourse-level phenomena and discuss issues in the context usage abilities of context-aware MT models.

## 4.2 Related Work

**Context Usage in Language Generation** An appropriate<sup>2</sup> usage of input information is fundamental in tasks such as summarization (Maynez et al., 2020) to ensure the soundness of generated texts. While appropriateness is traditionally verified post-hoc using trained models (Durmus et al., 2020; Kryscinski et al., 2020; Goyal and Durrett, 2021), recent interpretability works aim to gauge input influence on model predictions using internal properties of language models, such as the mixing of contextual information across model layers (Kobayashi et al., 2020; Ferrando et al., 2022b; Mohebbi et al., 2023) or the layer-by-layer refinement of next token predictions (Geva et al., 2022; Belrose et al., 2023). Recent attribution methods can disentangle factors influencing generation in language models (Yin and Neubig, 2022) and were successfully used to detect and mitigate hallucinatory behaviors (Tang et al., 2022; Dale et al., 2023a,b). Our proposed method adopts this intrinsic perspective to identify context reliance without ad hoc trained components.

**Context Usage in Neural Machine Translation** Despite advances in context-aware MT (Voita et al., 2018; Voita et al., 2019a; Lopes et al., 2020; Majumder et al., 2022; Jin et al., 2023; *inter alia*, surveyed by Maruf et al., 2021), only a few works explored whether context usage in MT models aligns with human intuition. Notably, some studies focused on *which parts of context* inform model predictions, finding that supposedly

---

<sup>1</sup>Code: <https://github.com/gsarti/pecore>

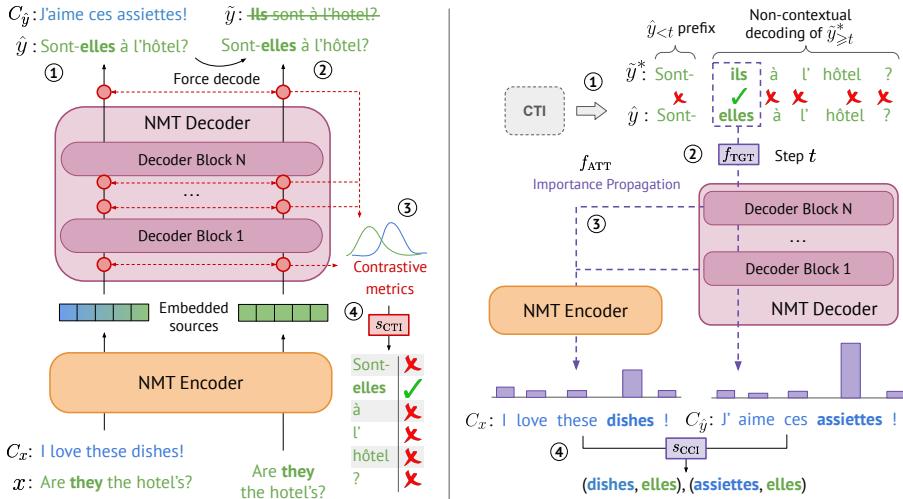
<sup>2</sup>We avoid using the term *faithfulness* due to its ambiguous usage in interpretability research.

context-aware MT models are often incapable of using contextual information (Kim et al., 2019b; Fernandes et al., 2021) and tend to pay attention to irrelevant words (Voita et al., 2018), with an overall poor agreement between human annotations and model rationales (Yin et al., 2021). Other works instead investigated *which parts of generated texts* are influenced by context, proposing various contrastive methods to detect gender biases, over- and under-translations (Vamvas and Sennrich, 2021a, 2022), and to identify various discourse-level phenomena in MT corpora (Fernandes et al., 2023b). While these two directions have generally been investigated separately, our work proposes a unified framework to enable an end-to-end evaluation of context-reliance plausibility in language models.

**Plausibility evaluation in NLP** Plausibility evaluation for NLP models has primarily focused on classification models (DeYoung et al., 2020; Atanasova et al., 2020; Attanasio et al., 2023). While few works investigate plausibility in language generation (Vafa et al., 2021; Ferrando et al., 2023), such evaluations typically involve a single generation step to complete a target sentence with a token connected to preceding information (e.g. subject/verb agreement, as in “*The pictures of the cat [is/are]*”), effectively biasing the evaluation by using a pre-selected token of interest. On the contrary, our framework proposes a more comprehensive evaluation of generation plausibility that includes the identification of context-sensitive generated tokens as an important prerequisite.

### 4.3 The PECoRE Framework

PECoRE is a two-step framework for identifying context dependence in generative language models. First, *context-sensitive tokens identification* (CTI) selects which tokens among those generated by the model were influenced by the presence of the preceding context (e.g. the feminine options “alla pastorella, le” in Figure 4.1). Then, *contextual cues imputation* (CCI) attributes the prediction of context-sensitive tokens to specific cues in the provided context (e.g. the feminine cues “she, Le pecore” in Figure 4.1). **Cue-target pairs** formed by influenced target tokens and their respective influential context cues can then be compared to human rationales to assess the models’ plausibility of context reliance for contextual phenomena of interest. Figure 4.2 provides an overview of the two steps applied to the context-aware MT setting discussed by this work. A more general formalization of the framework for language generation is proposed in the following sections.



**Figure 4.2 |** The PECoRe framework applied to an encoder-decoder MT model. **Left:** Context-sensitive token identification (CTI). ①: A context-aware MT model translates source context ( $C_x$ ) and current ( $x$ ) sentences into target context ( $C_{\hat{y}}$ ) and current ( $\hat{y}$ ) outputs. ②:  $\hat{y}$  is force-decoded in the non-contextual setting instead of natural output  $\tilde{y}$ . ③: Contrastive metrics are collected throughout the model for every  $\hat{y}_i$  token to compare the two settings. ④: Selector  $s_{cti}$  maps metrics to binary context-sensitive labels for every  $\hat{y}_i$ . **Right:** Contextual cues imputation (CCI). ①: Non-contextual target  $\tilde{y}^*$  is generated from contextual prefix  $\hat{y}_{<t}$ . ②: Function  $f_{tgt}$  is selected to contrast model predictions with  $(\hat{y}_t)$  and without  $(\tilde{y}_t^*)$  input context. ③: Attribution method  $f_{att}$  using  $f_{tgt}$  as target scores contextual cues driving  $\hat{y}_t$  prediction. ④: Selector  $s_{cci}$  selects relevant cues, and cue-target pairs are assembled.

### 4.3.1 Notation

Let  $X_{ctx}^i$  be the sequence of contextual inputs containing  $N$  tokens from vocabulary  $\mathcal{V}$ , composed by current input  $x$ , generation prefix  $y_{<i}$  and context  $C$ . Let  $X_{no-ctx}^i$  be the non-contextual input in which  $C$  tokens are excluded.<sup>3</sup>  $P_{ctx}^i = P(x, y_{<i}, C, \theta)$  is the discrete probability distribution over  $\mathcal{V}$  at generation step  $i$  of a language model with  $\theta$  parameters receiving contextual inputs  $X_{ctx}^i$ . Similarly,  $P_{no-ctx}^i = P(x, y_{<i}, \theta)$  is the distribution obtained from the same model for non-contextual input  $X_{no-ctx}^i$ . Both distributions are equivalent to vectors in the probability simplex in  $\mathbb{R}^{|\mathcal{V}|}$ , and we use  $P_{ctx}(y_i)$  to denote the probability of next token  $y_i$  in  $P_{ctx}^i$ , i.e.  $P(y_i | x, y_{<i}, C)$ .

<sup>3</sup>In the contextual MT example of Figure 4.2,  $C$  includes source context  $C_x$  and target context  $C_y$ .

### 4.3.2 Context-sensitive Token Identification (CTI)

CTI adapts the contrastive conditioning paradigm proposed by Vamvas and Sennrich (2021a) to detect input context influence on model predictions using the contrastive pair  $P_{\text{ctx}}^i, P_{\text{no-ctx}}^i$ . Both distributions are relative to the **contextual target sentence**  $\hat{y} = \{\hat{y}_1 \dots \hat{y}_n\}$ , corresponding to the sequence produced by a decoding strategy of choice in the presence of input context. In Figure 4.2, the contextual target sentence  $\hat{y} = \text{"Sont-elles à l'hôtel?"}$  is generated when  $x$  and contexts  $C_x, C_{\hat{y}}$  are provided as inputs, while **non-contextual target sentence**  $\tilde{y} = \text{"Ils sont à l'hôtel?"}$  would be produced when only  $x$  is provided. In the latter case,  $\hat{y}$  is instead force-decoded from the non-contextual setting to enable a direct comparison of matching outputs. We define a set of **contrastive metrics**  $\mathcal{M} = \{m_1, \dots, m_M\}$ , where each  $m : \Delta_{|\mathcal{V}|} \times \Delta_{|\mathcal{V}|} \mapsto \mathbb{R}$  maps a contrastive pair of probability vectors to a continuous score. For example, the difference in next token probabilities for contextual and non-contextual settings, i.e.  $P_{\text{diff}}(\hat{y}_i) = P_{\text{ctx}}(\hat{y}_i) - P_{\text{no-ctx}}(\hat{y}_i)$ , might be used for this purpose.<sup>4</sup> Target tokens with high contrastive metric scores can be identified as *context-sensitive*, provided  $C$  is the only added parameter in the contextual setting. Finally, a **selector** function  $s_{\text{cti}} : \mathbb{R}^{|\mathcal{M}|} \mapsto \{0, 1\}$  (e.g. a statistical threshold selecting salient scores) is used to classify every  $\hat{y}_i$  as context-sensitive or not.

### 4.3.3 Contextual Cues Imputation (CCI)

CCI applies the contrastive attribution paradigm (Yin and Neubig, 2022) to trace the generation of every context-sensitive token in  $\hat{y}$  back to the context  $C$ , identifying the cues that drive model predictions.

**Definition 4.3.1.** Let  $s, s'$  be the resulting scores of two attribution target functions  $f_{\text{tgt}}, f'_{\text{tgt}}$ . An attribution method  $f_{\text{att}}$  is **target-dependent** if importance scores  $A$  are computed in relation to the outcome of its attribution target function, i.e. whenever the following condition is verified.

$$f_{\text{att}}(x, y_{<t}, C, \theta, s) \neq f_{\text{att}}(x, y_{<t}, C, \theta, s') \quad \forall s \neq s'$$

In practice, common gradient-based attribution approaches (Simonyan et al., 2014; Sundararajan et al., 2017) are target-dependent as they rely on the outcome predicted by the model (typically the logit or the probability of the predicted class) as the differentiation target to backpropagate importance to model input features. Similarly,

---

<sup>4</sup>We use  $m^i$  to denote the result of  $m(P_{\text{ctx}}^i, P_{\text{no-ctx}}^i)$ . Several metrics are presented in Section 4.4.2.

perturbation-based approaches (Zeiler and Fergus, 2014) use the variation in prediction probability for the predicted class when noise is added to some of the model inputs to quantify the importance of the noised features.

On the contrary, recent approaches that rely solely on model internals to define input importance are generally *target-insensitive*. For example, attention weights used as model rationales, either in their raw form or after a rollout procedure to obtain a unified score (Abnar and Zuidema, 2020), are independent of the predicted outcome. Similarly, value zeroing scores (Mohebbi et al., 2023) reflect only the representational dissimilarity across model layers before and after zeroing value vectors, and as such do not explicitly account for model predictions.

**Definition 4.3.2.** Let  $\mathcal{T}$  be the set of indices corresponding to context-sensitive tokens identified by the CTI step, such that  $t \in \hat{y}$  and  $\forall t \in \mathcal{T}, s_{\text{cti}}(m_1^t, \dots, m_M^t) = 1$ . Let also  $f_{\text{tgt}} : \Delta_{|\mathcal{V}|} \times \dots \mapsto \mathbb{R}$  be a **contrastive attribution target** function representing an attribution target of interest, for example, the difference in next-token probabilities between the contextual option  $\hat{y}_t$  and the non-contextual option  $\tilde{y}_t^*$  from the same contextual distribution  $P_{\text{ctx}}^t$ , plus any additional required parameter. The **contrastive attribution method**  $f_{\text{att}}$  is a composite function quantifying the importance of contextual inputs to determine the output of  $f_{\text{tgt}}$  for a given model with  $\theta$  parameters.

$$f_{\text{att}}(\hat{y}_t) = f_{\text{att}}(x, \hat{y}_{<t}, C, \theta, f_{\text{tgt}}) = f_{\text{att}}(x, \hat{y}_{<t}, C, \theta, f_{\text{tgt}}(P_{\text{ctx}}^t, \dots))$$

**Remark 4.3.1.** The non-contextual next token  $\tilde{y}_t^*$  can be computed using the contextual prefix  $\hat{y}_{<t} = \{\hat{y}_1, \dots, \hat{y}_{t-1}\}$  (e.g.  $\hat{y}_{<t}$  = “Sont-” in Figure 4.2) and non-contextual inputs  $X_{\text{no-ctx}}^t$ . This is conceptually equivalent to predicting the next token of a new non-contextual sequence  $\tilde{y}^*$  which, contrary to the original  $\tilde{y}$ , starts from a forced contextual prefix  $\hat{y}_{<t}$  (e.g. “ils” in  $\tilde{y}^*$  = “Sont-ils à l'hôtel?” in Figure 4.2).

**Remark 4.3.2.** A  $f_{\text{tgt}}$  making use of both  $P_{\text{ctx}}^t$  and  $P_{\text{no-ctx}}^t$ , e.g. the KL divergence between the contextual and non-contextual probability distributions (Kullback and Leibler, 1951), can ultimately result in non-zero  $f_{\text{att}}(\hat{y}_t)$  scores, even when  $\hat{y}_t = \tilde{y}_t^*$ , i.e. even when the next predicted token is the same, since probabilities  $P_{\text{ctx}}(\hat{y}_t), P_{\text{no-ctx}}(\tilde{y}_t^*)$  are likely to differ beyond top-1 predictions. This is a desirable property of  $f_{\text{att}}$ , as it allows the attribution method to capture the influence of context on the model’s decision-making process, even in the case where the predicted token remains unchanged.

**Remark 4.3.3.** Our formalization of  $f_{\text{att}}$  generalizes the method proposed by Yin and Neubig (2022) to support any target-dependent attribution method, such as popular

gradient-based approaches (Simonyan et al., 2014; Sundararajan et al., 2017), and any contrastive attribution target  $f_{\text{tgt}}$ .

$f_{\text{att}}$  produces a sequence of attribution scores  $A_t = \{a_1, \dots, a_N\}$  matching contextual input length  $N$ . From those, only the subset  $A_{t \text{ ctx}}$  of scores corresponding to context input sequence  $C$  are passed to **selector** function  $s_{\text{cci}} : \mathbb{R} \mapsto \{0, 1\}$ , which predicts a set  $\mathcal{C}_t$  of indices corresponding to contextual cues identified by CCI, such that  $\forall c \in \mathcal{C}_t, \forall a \in A_{t \text{ ctx}}, s_{\text{cci}}(a_c) = 1$ .

Having collected all context-sensitive generated token indices  $\mathcal{T}$  using CTI and their contextual cues through CCI ( $\mathcal{C}_t$ ), PECoRE ultimately returns a sequence  $S_{\text{ct}}$  of all identified cue-target pairs:

$$\begin{aligned}\mathcal{T} &= \text{CTI}(C, x, \hat{y}, \theta, \mathcal{M}, s_{\text{cti}}) = \{t \mid s_{\text{cti}}(m_1^t, \dots, m_M^t) = 1\} \\ \mathcal{C} &= \text{CCI}(\mathcal{T}, C, x, \hat{y}, \theta, f_{\text{att}}, f_{\text{tgt}}, s_{\text{cci}}) = \{c \mid s_{\text{cci}}(a_c) = 1 \forall a_c \in A_{t \text{ ctx}}, \forall t \in \mathcal{T}\} \\ S &= \text{PECoRE}(C, x, \theta, s_{\text{cti}}, s_{\text{cci}}, \mathcal{M}, f_{\text{att}}, f_{\text{tgt}}) = \{(C_c, \hat{y}_t) \mid \forall t \in \mathcal{T}, \forall c \in \mathcal{C}_t, \forall \mathcal{C}_t \in \mathcal{C}\}\end{aligned}$$

A pseudocode implementation for the PECoRE algorithm is provided in Algorithm 1.

**Algorithm 1** PECoRe cue-target extraction process

---

**Require:**  $C, x$  (Input context and current sequences),  $\theta$  (Model parameters),  $s_{\text{cti}}, s_{\text{ccci}}$  (Selector functions),  $\mathcal{M}$  (Contrastive metrics),  $f_{\text{att}}$  (Contrastive attribution method),  $f_{\text{tgt}}$  (Contrastive attribution target function)

- 1: **procedure** PECoRe( $C, x, \theta, s_{\text{cti}}, s_{\text{ccci}}, \mathcal{M}, f_{\text{att}}, f_{\text{tgt}}$ )
- 2:    $\hat{y}$  = generate( $C, x, \theta$ ) using any decoding strategy and parameters
- 3:    $\mathcal{T} = \text{CTI}(C, x, \hat{y}, \theta, \mathcal{M}, s_{\text{cti}})$
- 4:   **for all**  $t \in \mathcal{T}$  **do**
- 5:      $\mathcal{C}_t = \text{CCI}(t, C, x, \hat{y}, \theta, f_{\text{att}}, f_{\text{tgt}}, s_{\text{ccci}})$
- 6:     **for all**  $c \in \mathcal{C}_t$  **do**
- 7:       Store  $(C_t^c, \hat{y}_t)$  in  $S_{\text{ct}}$
- 8:     **end for**
- 9:   **end for**
- 10:   **return**  $S_{\text{ct}}$  // Set of cue-target pairs
- 11: **end procedure**
- 12: **procedure** CTI( $C, x, \hat{y}, \theta, \mathcal{M}, s_{\text{cti}}$ )
- 13:    $\mathcal{T} = \emptyset$  // Empty set for context-sensitive indices of  $\hat{y}$  tokens
- 14:   **for all**  $\hat{y}_i \in \hat{y}$  **do**
- 15:     **for all**  $m \in \mathcal{M}$  **do**
- 16:        $m^i = m(P_{\text{ctx}}(\hat{y}_i), P_{\text{no-ctx}}(\hat{y}_i))$
- 17:     **end for**
- 18:     **if**  $(s_{\text{cti}}(m_1^i, \dots, m_M^i) = 1)$  **then**
- 19:       Store  $i$  in set  $\mathcal{T}$
- 20:     **end if**
- 21:   **end for**
- 22:   **return**  $\mathcal{T}$
- 23: **end procedure**
- 24: **procedure** CCI( $t, C, x, \hat{y}, \theta, f_{\text{att}}, f_{\text{tgt}}, s_{\text{ccci}}$ )
- 25:    $\mathcal{C}_t = \emptyset$  // Empty set for contextual cues for target token  $t$
- 26:   Generate constrained non-contextual target current sequence  $\hat{y}^*$  from  $\hat{y}_{<t}$
- 27:   Use attribution method  $f_{\text{att}}$  with target  $f_{\text{tgt}}$  to get importance scores  $A_t$
- 28:   Identify the subset  $A_{t \text{ ctx}}$  corresponding to tokens of context  $C = \{C_1, \dots, C_K\}$
- 29:   **for all**  $a_i \in A_{t \text{ ctx}} = \{a_1, \dots, a_K\}$  **do**
- 30:     **if**  $s_{\text{ccci}}(a_i) = 1$  **then**
- 31:       Store  $C_i$  in  $\mathcal{C}_t$
- 32:     **end if**
- 33:   **end for**
- 34:   **return**  $\mathcal{C}_t$
- 35: **end procedure**

---

## 4.4 Context Reliance Plausibility in Context-aware MT

This section describes our evaluation of PECoRE in a controlled setup. We experiment with several contrastive metrics and attribution methods for CTI and CCI (Section 4.4.2, Section 4.4.5), evaluating them in isolation to quantify the performance of individual components. An end-to-end evaluation is also performed in Section 4.4.5 to establish the applicability of PECoRE in a naturalistic setting.

### 4.4.1 Experimental Setup

**Evaluation Datasets** Evaluating generation plausibility requires human annotations for context-sensitive tokens in target sentences and disambiguating cues in their preceding context. To our knowledge, the SCAT dataset (Yin et al., 2021) is the only resource matching these requirements. SCAT is an English→French corpus with human annotations of anaphoric pronouns and disambiguating context on OpenSubtitles2018 dialogue translations (Lison et al., 2018; Lopes et al., 2020). SCAT examples were extracted automatically using lexical heuristics and thus contain only a limited set of anaphoric pronouns (*it, they* → *il/elle, ils/elles*), with no guarantees of contextual cues being found in preceding context.

The original SCAT test set contains 1000 examples with automatically identified context-sensitive pronouns *it/they* (marked by `<p>...</p>`) and human-annotated contextual cues aiding their disambiguation (marked by `<hon>...</hoff>`). Of these, we find 38 examples containing malformed tags and several more examples where an unrelated word containing *it* or *they* was wrongly marked as context-sensitive (e.g. `the soccer ball h<p>it</p> your chest`). Moreover, due to the original extraction process adopted for SCAT, there is no guarantee that contextual cues will be contained in the preceding context, as they could also appear in the same sentence, defeating the purpose of our context usage evaluation. Thus, we prefilter the entire corpus to retain only sentences with well-formed tags and inter-sentential contextual cues identified by the original annotators. Moreover, a manual inspection procedure is carried out to validate the original cue tags and discard problematic sentences, obtaining a final set of 250 examples with inter-sentential pronoun coreference, which we name SCAT+<sup>5</sup>.

Additionally, we manually annotate contextual cues in DiscEval-MT (Bawden et al., 2018), another English→French corpus containing handcrafted examples for *anaphora resolution* (ANA) and *lexical choice* (LEX). In the case of DiscEval-MT, we use minimal

---

<sup>5</sup>SCAT+ is available on the Hugging Face Hub: [inseq/scat](#)

pairs in the original dataset to automatically mark differing tokens as context-sensitive. Then, contextual cues are manually labeled separately by two annotators with good familiarity with both English and French. Cue annotations are compared across the two splits, resulting in very high agreement due the simplicity of the corpus (97% overlap for ANA, 90% for LEX).<sup>6</sup>

Our final evaluation set contains 250 SCAT+ and 400 DiscEVAL-MT translations across two discourse phenomena. Table 4.1 provides some examples for the three data splits.

**Models** We evaluate two bilingual Opus models (Tiedemann and Thottingal, 2020) using the transformer base architecture (Vaswani et al., 2017, Small and Large), and mBART-50 1-to-many (Tang et al., 2021), a larger multilingual MT model supporting 50 target languages, using the `transformers` library (Wolf et al., 2020). We fine-tune models using extended translation units (Tiedemann and Scherrer, 2017) with contextual inputs marked by break tags such as `[source context <brk> source current]` to produce translations in the format `target context <brk> target current`, where context and current target sentences are generated. We perform context-aware fine-tuning on 242k IWSLT 2017 English→French examples (Cettolo et al., 2017), using a dynamic context size of 0–4 preceding sentences to ensure robustness to different context lengths and allow contextless usage. To further improve models’ context sensitivity, we continue fine-tuning on the SCAT training split, containing 11k examples with inter- and intra-sentential pronoun anaphora.

**Model Disambiguation Accuracy** We estimate contextual disambiguation accuracy by verifying whether annotated (gold) context-sensitive words are found in model outputs. Results before and after context-aware fine-tuning are shown in Table 4.2. We find that fine-tuning improves translation quality and disambiguation accuracy across all tested models, with larger gains for anaphora resolution datasets that closely match the fine-tuning data. To gain further insight into these results, we use context-aware models to translate examples with and without context and identify a subset of *context-sensitive translations* (OK-CS) for which the correct target word is generated only when input context is provided to the model. Interestingly, we find a non-negligible amount of translations that are correctly disambiguated even in the absence of input context (corresponding to OK minus OK-CS in Table 4.2). For these examples, the correct prediction of ambiguous words aligns with model biases, such as defaulting to masculine gender for anaphoric pronouns (Stanovsky et al., 2019) or using the most frequent sense

---

<sup>6</sup>Our modified version of DiscEVAL-MT is available on the Hugging Face Hub: [inseq/disc\\_eval\\_mt](https://huggingface.co/inseq/disc_eval_mt).

---

**SCAT+**

*C<sub>x</sub>* : I loathe that song. But why did you bite poor Birdie's head off? Because I've heard it more times than I care to. It haunts me. Just stop, for a moment.

*C<sub>y</sub>* : Je hais cette chanson (song, FEMININE). Mais pourquoi avoir parlé ainsi à la pauvre Birdie ? Parce que j'ai entendu ce chant plus que de fois que je ne le peux. Elle (she) me hante. Arrêtez-vous un moment.

*x* : How does it haunt you?

*y* : Comment peut-elle (she) vous hanter?

---

*C<sub>x</sub>* : - Ah! Sven! It's been so long. - Riley, it's good to see you. - You, too. How's the boat? Uh, it creaks, it groans.

*C<sub>y</sub>* : Sven ! - Riley, contente de te voir. - Content aussi. Comment va le bateau (boat, MASCULINE)? Il (he) craque de partout.

*x* : Not as fast as it used to be.

*y* : Il (he) n'est pas aussi rapide qu'avant.

---

**DISCEVAL-MT ANA**

*C<sub>x</sub>* : But how do you know the woman isn't going to turn out like all the others?

*C<sub>y</sub>* : Mais comment tu sais que la femme (woman, FEMININE) ne finira pas comme toutes les autres?

*x* : This one's different.

*y* : Celle-ci (This one, FEMININE) est différente.

---

*C<sub>x</sub>* : Can you authenticate these signatures, please?

*C<sub>y</sub>* : Pourriez-vous authentifier ces signatures (FEMININE), s'il vous plaît?

*x* : Yes, they're mines.

*y* : Oui, ce sont les miennes (mines, FEMININE).

---

**DISCEVAL-MT LEX**

*C<sub>x</sub>* : Do you think you can shoot it from here?

*C<sub>y</sub>* : Tu penses que tu peux le tirer (shoot) dessus à partir d'ici?

*x* : Hand me that bow.

*y* : Passe-moi cet arc (bow, WEAPON).

---

*C<sub>x</sub>* : Can I help you with the wrapping?

*C<sub>y</sub>* : Est-ce que je peux t'aider pour l'emballage (wrapping)?

*x* : Hand me that bow.

*y* : Passe-moi ce ruban (bow, GIFT WRAP).

---

**Table 4.1 |** Examples from the SCAT+ and DISCEVAL-MT datasets used in our analysis with highlighted context-sensitive tokens and contextual cues used for plausibility evaluation using PECoRE. Glosses are added for French words of interest to facilitate understanding.

Model	SCAT+			DiscEval-MT (ANA)			DiscEval-MT (LEX)		
	BLEU	OK	OK-CS	BLEU	OK	OK-CS	BLEU	OK	OK-CS
Opus Small ( <i>def.</i> )	29.1	0.14	-	43.9	0.40	-	30.5	0.29	-
Opus Small S+T <sub>ctx</sub>	<u>39.1</u>	<u>0.81</u>	0.59	<u>48.1</u>	<u>0.60</u>	0.24	<u>33.5</u>	<u>0.36</u>	0.07
Opus Large ( <i>def.</i> )	29.0	0.16	-	39.2	0.41	-	31.2	0.31	-
Opus Large S+T <sub>ctx</sub>	<u>40.3</u>	<u>0.83</u>	0.58	<u>48.9</u>	<u>0.68</u>	0.31	<u>34.8</u>	<u>0.38</u>	0.10
mBART-50 ( <i>def.</i> )	23.8	0.26	-	33.4	0.42	-	24.5	0.25	-
mBART-50 S+T <sub>ctx</sub>	37.6	0.82	0.55	<u>49.0</u>	<u>0.62</u>	0.32	29.3	0.30	0.07

**Table 4.2 |** Translation quality of English→French MT models before (*def.*) and after (*S+T<sub>ctx</sub>*) context-aware MT fine-tuning. **OK:** % of translations with correct disambiguation for discourse phenomena. **OK-CS:** % of translations where the correct disambiguation is achieved only when context is provided.

for word sense disambiguation. Provided that such examples are unlikely to exhibit context reliance, we focus particularly on the OK-CS subset results in our following evaluation.

#### 4.4.2 Metrics for Context-sensitive Target Identification

The following contrastive metrics are evaluated for detecting context-sensitive tokens in the CTI step.

**Relative Context Saliency** We use contrastive gradient norm attribution (Yin and Neubig, 2022) to compute input importance towards predicting the next token  $\hat{y}_i$  with and without input context. Positive importance scores are obtained for every input token using the L2 gradient vectors norm (Bastings et al., 2022), and relative context saliency is obtained as the proportion between the normalized importance for context tokens  $c \in C_x, C_y$  and the overall input importance, following previous work quantifying MT input contributions (Voita et al., 2021; Ferrando et al., 2022a; Edman et al., 2024).

$$\nabla_{\text{ctx}}(P_{\text{ctx}}^i, P_{\text{no-ctx}}^i) = \frac{\sum_{c \in C_x, C_y} \|\nabla_c(P_{\text{ctx}}(\hat{y}_i) - P_{\text{no-ctx}}(\hat{y}_i))\|}{\sum_{t \in X_{\text{ctx}}^i} \|\nabla_t(P_{\text{ctx}}(\hat{y}_i) - P_{\text{no-ctx}}(\hat{y}_i))\|}$$

**Likelihood Ratio (LR) and Pointwise Contextual Cross-mutual Information (P-CXMI)** Proposed by Vamvas and Sennrich (2021a) and Fernandes et al. (2023b), respectively, both metrics frame context dependence as a ratio of contextual and non-contextual probabilities.

$$\text{LR}(P_{\text{ctx}}^i, P_{\text{no-ctx}}^i) = \frac{P_{\text{ctx}}(\hat{y}_i)}{P_{\text{ctx}}(\hat{y}_i) + P_{\text{no-ctx}}(\hat{y}_i)}$$

$$\text{P-CXMI}(P_{\text{ctx}}^i, P_{\text{no-ctx}}^i) = -\log \frac{P_{\text{ctx}}(\hat{y}_i)}{P_{\text{no-ctx}}(\hat{y}_i)}$$

**KL-Divergence** (Kullback and Leibler, 1951) between  $P_{\text{ctx}}^i$  and  $P_{\text{no-ctx}}^i$  is the only metric we evaluate that considers the full distribution rather than the probability of the predicted token. We include it to test the intuition that the impact of context inclusion might extend beyond top-1 token probabilities.

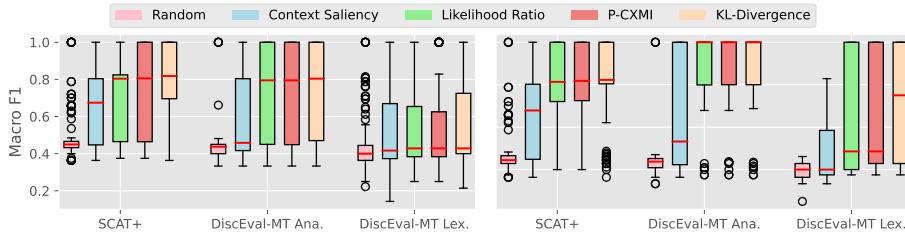
$$D_{\text{KL}}(P_{\text{ctx}}^i \| P_{\text{no-ctx}}^i) = \sum_{\hat{y}_i \in \mathcal{V}} P_{\text{ctx}}(\hat{y}_i) \log \frac{P_{\text{ctx}}(\hat{y}_i)}{P_{\text{no-ctx}}(\hat{y}_i)}$$

#### 4.4.3 Plausibility Evaluation Metrics

In practice, the CTI and CCI steps in PECoRE produce a sequence of continuous scores that are later binarized using selectors  $s_{\text{cti}}, s_{\text{cci}}$ , introduced in Section 4.3. To evaluate their validity, those are compared to a sequence  $I_h$  of the same length containing binary values, where 1s correspond to the cues identified by human annotators, while the rest are set to 0. In our experiments, we use two standard plausibility metrics introduced by DeYoung et al. (2020):

**Token-level Macro F1** is the harmonic mean of precision and recall at the token level, using  $I_h$  as the ground truth and the post-selector binarized scores as predictions. Macro-averaging is used to account for the sparsity of cues in  $I_h$ . We use this metric in our primary analysis, as the discretization step is more likely to reflect realistic plausibility performance, since it matches more closely the annotation process used to derive  $I_h$ . We note that Macro F1 can be considered a lower bound for plausibility, as the results depend heavily on the choice of the selector used for discretization.

**Area Under Precision-Recall Curve (AUPRC)** is computed as the area under the curve obtained by varying a threshold over token importance scores and computing the precision and recall for resulting discretized  $I_m$  predictions while keeping  $I_h$  as the ground truth. Contrary to Macro F1, AUPRC is selector-independent and accounts for tokens' relative ranking and degree of importance. Consequently, it acts as an upper bound for plausibility, as if the optimal selector was used. Results using AUPRC are presented in Section A.2.2 for completeness, but we focus on Macro F1 in the primary analysis.



**Figure 4.3 |** Macro F1 of contrastive metrics for context-sensitive target token identification (CTI) using Opus Large on the full datasets (left) or on ok-cs context-sensitive subsets (right).

#### 4.4.4 CTI Plausibility Results

Figure 4.3 presents our metrics evaluation for CTI, with results for the full test sets and the subsets of context-sensitive sentences (ok-cs) highlighted in Table 4.2. To keep our evaluation simple, we use a naive  $s_{cti}$  selector tagging all tokens with metric scores one standard deviation above the per-example mean as context-sensitive. We also include a stratified random baseline matching the frequency of occurrence of context-sensitive tokens in each dataset. Datapoints in Figure 4.3 are sentence-level macro F1 scores computed for every dataset example.

Pointwise metrics (LR, P-CXMI) show high plausibility for the context-sensitive subsets ok-cs across all datasets and models, but achieve lower performances on the full test set, especially for lexical choice phenomena less present in MT models' training. KL-Divergence performs on par with or better than pointwise metrics, suggesting that distributional shifts beyond top prediction candidates can provide helpful information for detecting context sensitivity. On the contrary, the poor performance of context saliency suggests that aggregate context reliance cannot reliably predict context sensitivity. A manual examination of misclassified examples reveals several context-sensitive tokens that were not annotated as such, as they did not match the dataset's phenomena of interest, but were still identified by CTI metrics. Table 4.3 presents several examples illustrating the contextual influence of French pronoun formality, whereas SCAT+ examples focus solely on gender disambiguation for anaphoric pronouns. This suggests that our evaluation of CTI metrics' plausibility can be considered a lower bound for actual method accuracy, as it is restricted to the two phenomena available in the datasets we used (anaphora resolution and lexical choice), rather than the broad set of contextual dependence phenomena. These results further underscore the importance of data-driven, end-to-end approaches like PECoRE in limiting the influence of selection bias

**Pronoun Grammatical Formality, SCAT+**

*C<sub>x</sub>* : [...] That demon that was in you, it wants you. But not like before. I think it loves you.  
*C<sub>y</sub>* : [...] Ce démon qui était en **vous**, il **vous** veut. Mais pas comme avant. Je pense qu'il **vous** aime.  
*x* : And it's powerless without you.  
*y* : Er il est impuissant sans **vous** (you, 2ND P. PLUR., FORMAL).

*C<sub>x</sub>* : You threaten my father again, I'll kill you myself... on this road. You hear me?  
*C<sub>y</sub>* : **Tu** menaces encore mon père, je **te** tuerai moi-même... sur cette route. **Tu** m'entends?  
*x* : Now it is with you as well.  
*y* : Maintenant elle est aussi avec **toi** (you, 2ND P. SING., INFORMAL).

*C<sub>x</sub>* : She went back to Delhi. What do you think? [...] Girls, I tell you.  
*C<sub>y</sub>* : Elle est revenue à Delhi. Qu'en penses-tu? [...] Les filles, je **te** le dis.  
*x* : I wish they were all like you.  
*y* : J'aimerais qu'elles soient toutes comme **toi** (you, 2ND P. SING., INFORMAL).

**Table 4.3 |** Examples of SCAT+ sentences with context-sensitive target tokens identified by CTI but not originally labeled as context-dependent in the dataset, since they do not match the gendered pronoun rule match used to create SCAT+. Relevant formality contextual cues are highlighted, and glosses are added for French words of interest to facilitate understanding.

during evaluation.

#### 4.4.5 Methods for Contextual Cues Imputation

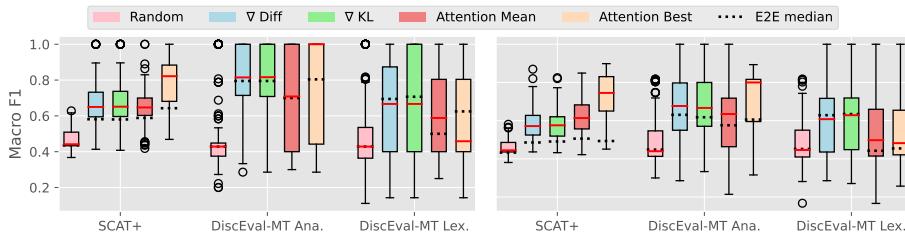
The following attribution methods are evaluated for detecting contextual cues in the CCI step.

**Contrastive Gradient Norm** (Yin and Neubig, 2022) estimates the input tokens' contributions towards predicting a target token, rather than a contrastive alternative. We use this method to explain the generation of context-sensitive tokens in the presence and absence of context.

$$A_{t \text{ ctx}} = \{ \| \nabla_c (f_{\text{tgt}}(P_{\text{ctx}}^i, \dots)) \| \mid \forall c \in C \}$$

For the choice of  $f_{\text{tgt}}$ , we evaluate both probability difference  $P_{\text{ctx}}(\hat{y}_i) - P_{\text{no-ctx}}(\hat{y}_i)$ , conceptually similar to the original formulation, and the KL-Divergence of contextual and non-contextual distributions  $D_{\text{KL}}(P_{\text{ctx}}^i \| P_{\text{no-ctx}}^i)$ . We use  $\nabla_{\text{diff}}$  and  $\nabla_{\text{KL}}$  to identify gradient norm attribution in the two settings.  $\nabla_{\text{KL}}$  scores can be seen as the contribution of input tokens towards the shift in probability distribution caused by the presence of input context.<sup>7</sup>

<sup>7</sup>Provided that  $P_{\text{no-ctx}}(\hat{y}_i)$  does not depend on context, the  $\nabla_{\text{KL}}$  gradient is functionally equivalent to the gradient for the cross-entropy function  $H(P_{\text{ctx}}, P_{\text{no-ctx}}) = - \sum_{\hat{y}_i \in \mathcal{V}} P_{\text{ctx}}(\hat{y}_i) \log P_{\text{no-ctx}}(\hat{y}_i)$ .



**Figure 4.4 |** Macro F1 of CCI methods over full datasets using Opus Large models trained with only source context (left) or with source+target context (right). Boxes and red median lines show CCI results based on gold context-sensitive tokens. Dotted bars show median CCI scores obtained from context-sensitive tokens identified by KL-Divergence during CTI (E2E settings).

**Attention Weights** Following previous work, we use the mean attention weight across all heads and layers (**Attention Mean**, Kim et al. (2019b)) and the weight for the head obtaining the highest plausibility per-dataset (**Attention Best**, Yin et al. (2021)) as importance measures for CCI. Attention Best can be seen as a best-case estimate of attention performance but is not a viable metric in real settings, provided that the best attention head to capture a phenomenon of interest is unknown beforehand. Since attention weights are model byproducts unaffected by predicted outputs, we use only attention scores for the contextual setting  $P_{\text{ctx}}^i$  and ignore the contextless alternative when using these metrics.

#### 4.4.6 CCI Plausibility Results

We conduct a controlled CCI evaluation using gold context-sensitive tokens as the starting point to attribute contextual cues. Provided that gold context-sensitive tokens are only available in annotated reference translations, a simple option when applying CCI to those would involve using references as model generations. However, this was shown to be problematic by previous research, as it would induce a *distributional discrepancy* in model predictions (Vamvas and Sennrich, 2021b). For this reason, we let the model generate a natural translation and instead try to align tags to this new sentence using the AWESOME aligner (Dou and Neubig, 2021) with LABSE multilingual embeddings (Feng et al., 2022). While this process is not guaranteed to always result in accurate tags, it provides a good approximation of gold CTI annotations for model generation, which is suitable for our assessment. This corresponds to the baseline plausibility evaluation described in Section 2.2.2, allowing us to evaluate attribution methods in isolation, assuming perfect identification of context-sensitive tokens. Figure 4.4 presents our

results. Scores in the right plot are relative to the context-aware Opus Large model of Section 4.4.4 using both source and target context. Instead, the left plot presents results for an alternative version of the same model that was fine-tuned using only the source context (i.e., translating  $C_x, x \rightarrow y$  without producing the target context  $C_y$ ). Source-only context was used in previous context-aware MT studies (Fernandes et al., 2022), and we include it in our analysis to assess how the presence of target context impacts model plausibility. We finally validate the end-to-end plausibility of PECoRE-detected pairs using context-sensitive tokens identified by the best CTI metric from Section 4.4.4 (KL-Divergence) as the starting point for CCI, and using a simple statistical selector equivalent to the one used for CTI evaluation.

First, contextual cues are more easily detected for the source-only model using all evaluated methods. This finding corroborates previous evidence highlighting how context usage issues might emerge when lengthy context is provided (Fernandes et al., 2021; Shi et al., 2023). When moving from gold CTI tags to the end-to-end setting (E2E) we observe a larger drop in plausibility for the SCAT+ and DiscEval-MT ANA datasets that more closely match the fine-tuning data of analyzed MT models. This suggests that standard evaluation practices may overestimate model plausibility for in-domain settings and that our proposed framework can effectively mitigate this issue. Interestingly, the Attention Best method suffers the most from end-to-end CCI application, while other approaches are more mildly affected. This can result from attention heads failing to generalize to other discourse-level phenomena at test time, providing further evidence of the limitations of attention as an explanatory metric (Jain and Wallace, 2019; Bastings and Filippova, 2020). While  $\nabla_{\text{diff}}$  and  $\nabla_{\text{KL}}$  appear as the most robust choices across the two datasets, per-example variability remains high across the board, leaving space for improvement for more plausible attribution methods in future work.

## 4.5 Detecting Context Reliance in the Wild

We continue our analysis by applying the PECoRE method to the popular Flores-101 MT benchmark (Goyal et al., 2022), containing groups of 3–5 contiguous sentences from English Wikipedia. While previous sections used labeled examples to evaluate the effectiveness of PECoRE components, here we apply our framework end-to-end to unannotated MT outputs and inspect the resulting cue-target pairs to identify the successes and failures of context-aware MT models.

Specifically, we apply PECoRE to the context-aware Opus Large and mBART-50 models of Section 4.4.1, using KL-Divergence as CTI metric and  $\nabla_{\text{KL}}$  as CCI attribution

**1. Acronym Translation (English → French, correct but more generic)**

$C_x$  : Across the United States of America, there are approximately 400,000 known cases of Multiple Sclerosis (MS) [...]

$C_y$  : Aux États-Unis, il y a environ 400 000 cas connus de sclérose en plaques [...]

$x$  : MS affects the central nervous system, which is made up of the brain, the spinal cord and the optic nerve.

$\hat{y}$  : La SEP affecte le système nerveux central, composé du cerveau, de la moelle épinière et du nerf optique.

$\hat{y}$  : La maladie affecte le système nerveux central, composé du cerveau, de la moelle épinière et du nerf optique.

**2. Anaphora Resolution (English → French, incorrect)**

$C_x$  : The terrified King and Madam Elizabeth were forced back to Paris by a mob of market women.

$C_y$  : Le roi et Madame Elizabeth ont été forcés à revenir à Paris par une foule de femmes du marché.

$x$  : In a carriage, they traveled back to Paris surrounded by a mob of people screaming and shouting threats [...]

$\hat{y}$  : Dans une carrière, ils sont retournés à Paris entourés d'une foule de gens hurlant et criant des menaces [...]

$\hat{y}$  : Dans une carrière, elles sont retournées à Paris entourées d'une foule de gens hurlant et criant des menaces [...]

**Table 4.4** | Flores-101 examples with cue-target pairs identified by PECoRE in Opus Large contextual translations. Context-sensitive tokens generated instead of their non-contextual counterparts are identified by CTI, and contextual cues justifying their predictions are retrieved by CCI. Other changes in  $\hat{y}$  are not considered context-sensitive by PECoRE.

method. We set  $s_{cti}$  and  $s_{cci}$  to two standard deviations above the per-example average score to focus our analysis on very salient tokens.

Table 4.4 and Table 4.5 show some examples annotated with PECoRE outputs. In the first example, the acronym MS, standing for Multiple Sclerosis, is translated generically as *la maladie* (the illness) in the contextual output, but as *SEP* (the French acronym for MS, i.e. *sclérose en plaques*) when context is not provided. PECoRE shows how this choice is mostly driven by the MS mention in source context  $C_x$  while the term *sclérose en plaques* in target context  $C_y$  is not identified as influential, possibly motivating the choice for the more generic option.

In the second example, the prediction of pronoun *elles* (they, feminine) depends on the context noun phrase *mob of market women* (*foule de femmes du marché* in French). However, the correct pronoun referent is *Le roi et Madame Elizabeth* (*the king and Madam Elizabeth*), so the pronoun should be the masculine default *ils*, commonly used for mixed-gender groups in French. PECoRE identifies this as a context-dependent failure due to an issue with the MT model’s anaphora resolution.

The third example presents an interesting case of erroneous numeric format cohesion that would typically go undetected when relying on pre-defined linguistic hypotheses. In this sentence, the score *26-00* is translated as *26* in the contextless output and as *26:00* in the context-aware translation. The *10:00* time indications found by PECoRE in the contexts suggest this is a case of problematic lexical cohesion.

---

**3. Numeric format cohesion (English → French, incorrect)**

---

$C_x$  : The games kicked off at 10:00 am with great weather apart from mid morning drizzle [...]

$C_y$  : Les matchs se sont éoulés à 10:00 du matin avec un beau temps à part la nuée du matin [...]

$x$  : South Africa started on the right note when they had a comfortable 26-00 win against Zambia.

$\hat{y}$  : L'Afrique du Sud a commencé sur la bonne note quand ils ont eu une confortable victoire de 26 contre le Zambia.

$\hat{y}$  : L'Afrique du Sud a commencé sur la bonne note quand ils ont eu une confortable victoire de 26:00 contre le Zambia.

---

**4. Lexical cohesion (English → Turkish, correct)**

---

$C_x$  : The activity of all stars in the system was found to be driven by their luminosity, their rotation, and nothing else.

$C_y$  : Sistemdeki bütün ıldızların faaliyetlerinin, parlaklıkları, rotasyonları ve başka hiçbir şeyin erkisi altında olduğunu ortaya çıkardılar.

$x$  : The luminosity and rotation are used together to determine a star's Rossby number, which is related to plasma flow.

$\hat{y}$  : Parlaklık ve döngü, bir yıldızın plazm akışıyla ilgili Rossby sayısını belirlemek için birlikte kullanılıyor.

$\hat{y}$  : Parlaklık ve rotasyon, bir ıldızın plazma akışıyla ilgili Rossby sayısını belirlemek için birlikte kullanılıyor.

---

**Table 4.5** | Flores-101 examples with cue-target pairs identified by PECoRE in mBART-50 contextual translations. Context-sensitive tokens generated instead of their non-contextual counterparts are identified by CTI, and contextual cues justifying their predictions are retrieved by CCI. Other changes in  $\hat{y}$  are not considered context-sensitive by PECoRE.

Finally, we include an example of context usage for English→Turkish translation to test the contextual capabilities of the default mBART-50 model without context-aware fine-tuning. Again, PECoRE shows how the word *rotasyon* (rotation) is selected over *döngü* (loop) as the correct translation in the contextual case due to the presence of the lexically similar word *rotasyonları* in the previous context.

## 4.6 Integrating PECoRE in Inseq

To facilitate the use of PECoRE in future research, a flexible implementation of the framework was incorporated into the Inseq toolkit presented in Chapter 3. Since its v0.6.0 Inseq offers the CLI command `attribute-context`, supporting all contrastive step functions and attribution methods in the library, and compatible with any decoder-only and encoder-decoder generative language model. Figure 4.5 provides an example employing the Inseq API to attribute a language model answer to input context paragraphs, similarly to the retrieval-augmented generation task we discuss in Chapter 5.<sup>8</sup> In

<sup>8</sup>The interface is available at: <https://huggingface.co/spaces/gsarti/pecore>.

**Input Context**

[1]: ① Hawaii consists of 137 volcanic islands that comprise almost the entire Hawaiian archipelago. The eight main islands, from northwest to southeast, are Ni'ihiu ②, Kaua'i, O'ahu, Moloka'i, Lāna'i, Kahoolawe, Maui, and Hawai'i, after which the state is named.

[2]: ① Of the 50 U.S. states, Hawaii is the eighth-smallest in land area and the 11th-least populous, but with 1.4 million residents, it ranks 13th in population density. Two-thirds of Hawaii residents live on O'ahu, home to the state's capital and largest city, Honolulu.

[3]: Historically dominated by a plantation economy, Hawaii remains a major agricultural exporter due to its fertile soil and uniquely tropical climate in the U.S. its economy has gradually diversified since the mid-20th century, with tourism and military defense becoming the two largest sectors. ①

**Query**

What is the Hawaii Island hosting most of the state residents?

**Generated Answer (Contextual)**

Based on the information provided, ① the Hawaii island is hosting the majority of the state residents. Two-thirds of Hawaii residents live on O'ahu, which is the largest and most populous island in the state. The other three islands (Ni'ihiu ②, Kaua'i, and Lāna'i) together make up the remaining 23% of Hawaii's population.

**Generated Answer (No Context)**

The island in the Hawaii archipelago that hosts most of the state residents is the island of O'ahu. O'ahu is the largest and most populous island in the Hawaiian Islands, [...] Other major islands in the state include Maui, Kauai, and the Big Island (also known as Hawaii Island).

**Figure 4.5 |** Example of context attribution for open-book QA using the Inseq-powered PECoRE demo. Context-sensitive tokens and contextual cues are highlighted.

In the example, the StableLM 2 Zephyr 1.6B language model<sup>9</sup> is prompted with contexts retrieved from Wikipedia to provide a long-form answer to a query about population in the Hawaiian islands. When referring to “the information provided” in ①, PECoRE identifies the indices of the two documents containing relevant information as salient. The name of Ni'ihiu, a small island with barely any population, is also found important when the model produces an additional remark on their population in ②. However, we observe that the answer in the context is not identified as salient by PECoRE during generation, suggesting that the model might be relying on memorization. We test the hypothesis by prompting the model in a closed-book setting without context paragraphs, finding that the model can indeed respond correctly without context. Moreover, as expected, the island of Ni'ihiu is never mentioned in the contextless response. Additional examples of PECoRE usage for other generation tasks are provided in Section A.2.3.

<sup>9</sup> [stabilityai/stablelm-2-zephyr-1\\_6b](https://github.com/stabilityai/stablelm-2-zephyr-1_6b)

#### 4.7 Conclusion

We introduced PECoRE, a novel interpretability framework for detecting and attributing context usage in language models' generations. PECoRE extends the standard plausibility evaluation procedure adopted in interpretability research by proposing a two-step procedure to identify context-sensitive generated tokens and match them to contextual cues contributing to their prediction. We applied PECoRE to context-aware MT, finding that context-sensitive tokens and their disambiguating rationales can be detected consistently and with reasonable accuracy across several datasets, models and discourse phenomena. Moreover, an end-to-end application of our framework without human annotations revealed incorrect context usage, leading to problematic MT model outputs.

While our evaluation is mainly focused on the machine translation domain, thanks to its generality and its integration in the Inseq framework PECoRE can easily be applied to other context-dependent language generation tasks such as question answering and summarization, as also demonstrated in the previous section. Future applications of our methodology could investigate the usage of in-context demonstrations and chain-of-thought reasoning in large language models (Brown et al., 2020; Wei et al., 2022), and explore PECoRE usage for different model architectures and input modalities. In the next chapter, we extend PECoRE for attributing context usage in retrieval-augmented generation tasks, where the model is expected to rely on external knowledge sources to produce answers to user queries.



# CHAPTER 5

---

## Answer Attribution for Trustworthy Retrieval-Augmented Generation

### Chapter Summary

This chapter extends our investigation of context usage to modern pre-trained decoder-only language models for the popular task of retrieval-augmented generation. We propose MIRAGE, an extension of PECoRE using context saliency to address trustworthiness challenges in answer attribution for RAG applications. We evaluate our proposed approach on a multilingual extractive QA dataset, finding high agreement with human-annotated answer attributions. On open-ended QA, MIRAGE achieves citation quality and efficiency comparable to self-citation prompting, while also allowing for a finer-grained control of attribution parameters. Our qualitative evaluation highlights the faithfulness of MIRAGE attributions and underscores the promising application of model internals for trustworthy answer generation with language models.

This chapter is adapted from the paper *Model Internals-based Answer Attribution for Trustworthy Retrieval-Augmented Generation* (Qi, Sarti, et al., 2024).

*“What makes the desert beautiful,” said the little prince, “is that it hides a well somewhere...”*

*– Antoine de Saint-Exupéry, Le petit prince (1943)*

## 5.1 Introduction

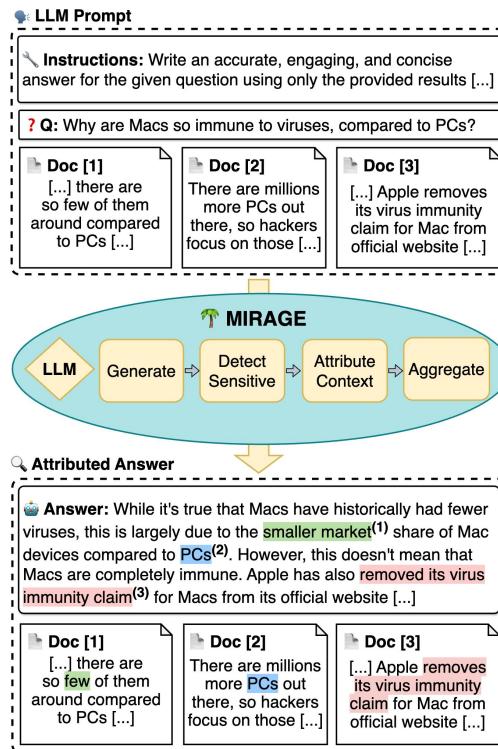
Retrieval-augmented generation with large language models has become the de facto standard methodology for question answering in both academic (Lewis et al., 2020; Izacard et al., 2023) and industrial settings (Dao and Le, 2023; Ma et al., 2024). This approach is effective in mitigating hallucinations and producing factually accurate answers (Petroni et al., 2020; Lewis et al., 2020; Borgeaud et al., 2022; Ren et al., 2025). However, verifying whether the model answer is faithfully supported by the retrieved sources is often non-trivial due to the large context size and the variety of potentially correct answers (Krishna et al., 2021; Xu et al., 2023a).

In light of this issue, several *answer attribution*<sup>1</sup> approaches were recently proposed to ensure the trustworthiness of RAG outputs (Rashkin et al., 2023; Bohnet et al., 2022; Muller et al., 2023). Initial efforts in this area employed models trained on Natural Language Inference (NLI) to automate the identification of supporting documents (Bohnert et al., 2022; Yue et al., 2023). However, being based on an external validator, this approach does not faithfully explain the answer generation process but simply identifies plausible sources supporting model answers in a post-hoc fashion. Following recent progress in the instruction-following abilities of LLMs, *self-citation* (i.e. prompting LLMs to generate inline citations alongside their answers) has been proposed to mitigate the training and inference costs of external validator modules (Gao et al., 2023a). However, self-citation performance is hindered by the imperfect instruction-following capacity of modern LLMs (Mu et al., 2023; Liu et al., 2023a), and resulting attributions are still predicted in an unintelligible, post-hoc fashion. This is an important limitation for these approaches, since the primary goal of answer attribution should be to ensure that the LLM is not ‘right for the wrong reasons’ (McCoy et al., 2019).

In light of these considerations, we introduce MIRAGE, an extension of the context-reliance evaluation PECoRE framework from the previous chapter for efficient and faithful answer attributions. MIRAGE first identifies context-sensitive tokens in a generated sentence by measuring the shift in LM predictive distribution caused by the added input context. Then, it attributes this shift to specific influential tokens in the context using gradient-based saliency or other input attribution techniques (Madsen et al., 2022a). Finally, attributions can be aggregated at the document level to match context-dependent generated sentences with retrieved documents that contribute to their prediction. The resulting pairs can then be converted to citations using the standard

---

<sup>1</sup>We use the term *answer attribution* (AA) when referring to the task of citing relevant sources to distinguish it from the *input attribution* methods used in MIRAGE.



**Figure 5.1 |** MIRAGE is a model internals-based framework for answer attribution in RAG settings. Context-sensitive answer spans (in color) are detected and matched with contextual cues in retrieved sources to evaluate the trustworthiness of models' answers.

answer attribution (AA) format.

We begin our assessment of MIRAGE on the short-form XOR-AttriQA dataset (Muller et al., 2023), showing high agreement between MIRAGE results and human annotations across several languages. We then test our method on the open-ended ELI5 dataset (Fan et al., 2019), achieving AA quality comparable to or better than self-citation while ensuring a higher degree of control over attribution parameters. In summary, we make the following contributions:<sup>2</sup>

- We introduce MIRAGE, a model internals-based answer attribution framework

<sup>2</sup>Code and data released at <https://github.com/Betswish/MIRAGE>. A demo for MIRAGE using the Inseq `attribute-context` API is available at <https://hf.co/spaces/gsarti/mirage>.

optimized for RAG applications.

- We quantify the plausibility of MIRAGE attributions on two datasets, showing improvements over NLI and self-citation methods while ensuring better controllability and efficiency.
- We analyze challenging attribution settings, highlighting MIRAGE’s faithfulness to LLMs’ reasoning process.

## 5.2 Background and Related Work

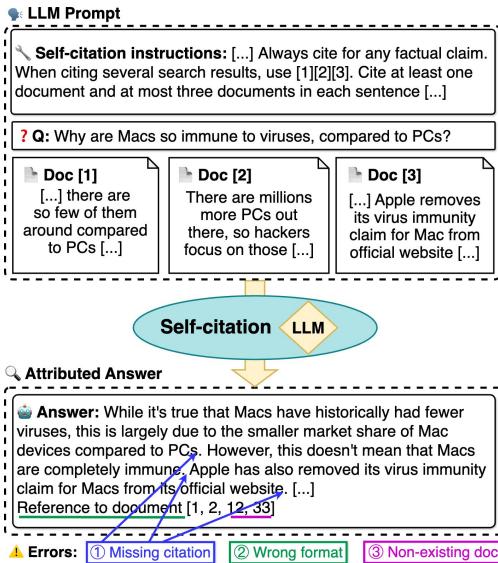
In RAG settings, a set of documents relevant to a user query is retrieved from an external dataset and infilled into an LLM prompt to improve the generation process (Petroni et al., 2020; Lewis et al., 2020). *Answer attribution* (Rashkin et al., 2023; Bohnet et al., 2022; Muller et al., 2023) aims to identify which retrieved documents support the generated answer (*answer faithfulness*, Gao et al., 2023b), e.g., by exploiting the similarity between model outputs and references.<sup>3</sup> Simplifying access to relevant sources via answer attribution is a fundamental step towards ensuring RAG trustworthiness in customer-facing scenarios (Liu et al., 2023a).

### 5.2.1 Answer Attribution Methods

**Entailment-based Answer Attribution** Bohnet et al. (2022) and Muller et al. (2023) propose to approximate human AA annotations with NLI systems such as TRUE (Honovich et al., 2022), using a source document as premise and an LLM-generated sentence as entailment hypothesis. AAs produced by these systems were shown to correlate strongly with human annotations, prompting their adoption in AA studies (Muller et al., 2023; Gao et al., 2023a). Despite their effectiveness, entailment-based methods can be computationally expensive when several answer sentence-document pairs are present. Moreover, this setup assumes the NLI model’s ability to robustly detect entailment relations across all domains and languages for which the LLM generator is used. In practice, however, NLI systems were shown to be brittle in challenging scenarios, exploiting shallow heuristics (McCoy et al., 2019; Nie et al., 2020; Sinha et al., 2021; Luo et al., 2022), and require dedicated efforts for less-resourced settings (Conneau et al., 2018). For example, NLI may fail to correctly attribute answers in multi-hop QA settings when considering individual documents as premises (Yang et al., 2018; Welbl et al., 2018).

---

<sup>3</sup>Popular frameworks such as `LangChain` and `LlamaIndex` support similarity-based citations using vector databases.



**Figure 5.2 |** Instruction-following errors in *self-citation*, using the setup of Gao et al. (2023a).

**Self-citation** (Gao et al., 2023a) is a recent AA approach exploiting the ability of recent LLMs to follow instructions in natural language (Raffel et al., 2020; Chung et al., 2024; Sanh et al., 2022; OpenAI, 2023), thereby avoiding the need for an external validator. Nakano et al. (2021) and Menick et al. (2022) propose citation fine-tuning for LLMs, while Gao et al. (2023a) instruct general-purpose LLMs to produce inline citations in a few-shot setting. Self-citation answers are generally more relevant to the provided sources' contents, but can still contain unsupported statements and inaccurate citations (Liu et al., 2023a). In our preliminary analysis, we find that self-citation often misses relevant citations, uses wrong formats, or refers to non-existing documents (Figure 5.2). For the ELI5 dataset (Fan et al., 2019), we find that LLaMA 2 7B Chat (Touvron et al., 2023) and Zephyr  $\beta$  7B (Tunstall et al., 2024) fail to produce AAs matching the prompt instructions for the majority of generated sentences, with almost all answers having at least one unattributed sentence when the Gao et al. (2023a) self-citation setup is used (Table 5.1).

**Answer Attribution can be Unfaithful** The aforementioned approaches do not account for attributions' *faithfulness*, i.e. whether the selected documents influence the

Model	Missing cit. (%)	
	Sent.	Ans.
Zephyr 7B $\beta$	54.5	95.7
LLaMA 2 7B	62.4	99.3

**Table 5.1 |** Percentage of unattributed sentences and answers with  $\geq 1$  unattributed sentences on ELI5.

LLM during the generation. Indeed, the presence of an entailment relation or high semantic similarity does not imply that the retrieved document influenced the answer generation process. This can be true in cases where LLMs may rely on memorized knowledge while ignoring relevant, albeit unnecessary, contextual information.

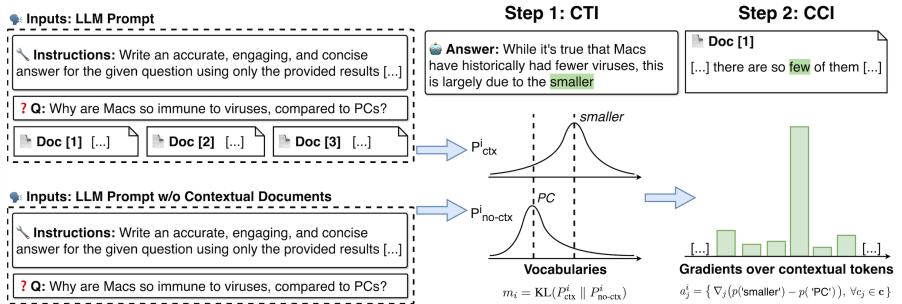
Even in the case of self-citation, recent work showed that, while the justifications of self-explaining LLMs appear plausible, they generally do not align with their internal reasoning process (Atanasova et al., 2023; Madsen et al., 2024; Agarwal et al., 2024; Randl et al., 2025), with little to no predictive efficacy (Huang et al., 2023). By contrast, approaches based on model internals are designed to faithfully reflect input importance in motivating model predictions. For instance, Alghisi et al. (2024) explores the use of gradient-based attribution to locate salient history segments for various dialogical tasks. Concurrent to our work, Phukan et al. (2024) and Cohen-Wang et al. (2024) have proposed other internals-based methods for granular AA of LLM generations. While the two-step approaches proposed in both works are similar to MIRAGE, they also differ in substantial ways. Notably, Phukan et al. (2024) derive attributions from embedding similarity, which does not capture the functional influence of context usage during the generation process. ContextCite (Cohen-Wang et al., 2024) instead fits a linear surrogate model to estimate the impact of ablating context segments on downstream answer probabilities. While this procedure approximates causal context influence, it still requires a sufficiently large context and many LLM forward passes to learn the surrogate model<sup>4</sup>, ultimately providing a coarser attribution for the full generated output. On the contrary, MIRAGE efficiently estimates generated tokens requiring attribution via contrastive metrics to produce granular attributions at the token level, limiting computations to estimate how context impacts LLM predictions. A maximally faithful AA approach would ablate all possible combinations of context elements to counterfactually estimate their importance in relation to model predictions. Given the long-form answers and contexts in RAG settings, this is practically unfeasible. Even if based on approximations, internals-based approaches such as MIRAGE are intrinsically more faithful than external validators like NLI models, since they aim to exploit information functional to the predictive process rather than relying solely on the generated output.

### 5.3 Method

Identifying which generated spans were most influenced by preceding information is a key challenge for LM attribution. The Model Internals-based RAG Explanations

---

<sup>4</sup>Authors suggest a minimum of 32 different ablations.



**Figure 5.3 |** Illustration of MIRAGE’s two-step approach adapted from PECoRE for RAG answer attribution. **Step 1:** CTI detects context-sensitive tokens in the generation (e.g. *smaller*). **Step 2:** CCI attributes the generation of detected tokens back to context tokens (e.g. *few* in Doc[1] promotes the generation of *smaller* instead of *PC*) using contrastive input attribution. Token pairs are then aggregated into sentence-document citations for practical usage.

(MIRAGE) method we propose is an extension of the Plausibility Evaluation for Context Reliance (PECoRE) framework (Sarti et al., 2024c) for context-aware machine translation. Importantly, this framework requires open-weights access to the LLM generator, which is a strict but necessary requirement to provide an accurate overview of the actual context usage during generation (Casper et al., 2024). This section frames the PECoRE’s two-step procedure in the context of RAG, as illustrated in Figure 5.3, and clarifies how MIRAGE adapts it for RAG answer attribution.

**Step 1: Context-sensitive Token Identification (CTI)** For every token in an answer sentence  $\mathbf{y} = \langle y_1, \dots, y_n \rangle$  generated by an LM prompted with a query  $\mathbf{q}$  and a context  $\mathbf{c} = \langle c_1, \dots, c_{|\mathbf{c}|} \rangle$ , a contrastive metric  $m$  such as KL divergence (Kullback and Leibler, 1951) is used to quantify the shift in the LM predictive distribution at the  $i$ -th generation step when the context is present or absent ( $P_{\text{ctx}}^i$  or  $P_{\text{no-ctx}}^i$ ). Resulting scores  $\mathbf{m} = \langle m_1, \dots, m_n \rangle$  reflect the context sensitivity of every generated token and can be converted into binary labels using a selector function  $s_{\text{cti}}$ :

$$\begin{aligned} \text{CTI}(\mathbf{q}, \mathbf{c}, \mathbf{y}) &= \{ y_i \mid s_{\text{cti}}(m_i) = 1 \forall y_i \in \mathbf{y} \} \\ \text{with } m_i &= D_{\text{KL}}(P_{\text{ctx}}^i \| P_{\text{no-ctx}}^i) \end{aligned}$$

**Step 2: Contextual Cues Imputation (CCI)** For every context-sensitive token  $y_i$  identified by CTI, a contrastive alternative  $y_i^{\setminus \mathbf{c}}$  is produced by excluding  $\mathbf{c}$  from the

prompt, but using the original generated prefix  $\mathbf{y}_{<i}$ . Then, *contrastive input attribution* (Yin and Neubig, 2022) is used to obtain attribution scores  $\mathbf{a}^i = \langle a_1^i, \dots, a_{|\mathbf{c}|}^i \rangle$  for every context token  $c_j \in \mathbf{c}$ :

$$\mathbf{a}^i = \{ \nabla_j (p(y_i) - p(y_i^{\setminus \mathbf{c}})), \forall c_j \in \mathbf{c} \}$$

where  $\nabla_j$  is the L2 norm of the gradient vector over the input embedding of context token  $c_j$ , and both probabilities are computed from the same contextual inputs  $(\mathbf{q}, \mathbf{c}, \mathbf{y}_{<i})$ . Intuitively, this procedure identifies which tokens in  $\mathbf{c}$  influence the increment of the probability for token  $y_i$  and the decrement of that for the non-contextual option  $y_i^{\setminus \mathbf{c}}$ , as shown in Step 2 in Figure 5.3. Resulting scores are once again binarized with a selector  $s_{\text{CCI}}$ :

$$\text{CCI}(y_i) = \{ c_j \mid s_{\text{cci}}(a_j^i) = 1, \forall c_j \in \mathbf{c} \}$$

This results in pairs of context-sensitive generated tokens and the respective input-context tokens influencing their prediction:

$$\mathcal{P} = \{ \langle y_i, c_j \rangle, \forall y_i \in \text{CTI}, \forall c_j \in \text{CCI}(y_i) \}$$

### 5.3.1 From Granular Attributions to Document-level Citations

**CTI Filtering** To obtain discrete labels from the CTI step, we set  $s_{\text{cti}}(m_i) = m_i \geq m^*$ , where  $m^*$  is a threshold value for selecting context-sensitive generated tokens. We experiment with two variants of  $m^*$ : a **calibrated threshold**  $m_{\text{cal}}^*$  obtained by maximizing agreement between the contrastive metric and human annotations on a calibration set with human AA annotations, and an **example-level threshold**  $m_{\text{ex}}^*$  using only within-example scores to avoid the need for calibration data. Following Sarti et al. (2024c), we set  $m_{\text{ex}}^* = \bar{\mathbf{m}} + \sigma_{\mathbf{m}}$ , where  $\bar{\mathbf{m}}$  and  $\sigma_{\mathbf{m}}$  are the average and standard deviation of  $\mathbf{m}$  scores for generated tokens.

**CCI Filtering** To extract granular document citations (i.e., colored spans with document indices in Figure 5.1), we set  $s_{\text{cci}} = a_j^i \geq a^{i*}$ , where  $a^{i*}$  is either the Top-K or Top-% highest attribution value in  $\mathbf{a}^i$ , to filter attributed context tokens  $c_j \in \text{CCI}(y_i)$ . Then, we use the identifier  $\text{docid}(c_j)$  of the documents they belong to as citation indices for context-sensitive token  $y_i$ . Highlights for consecutive tokens citing the same documents are collated into a single span and mapped from subword to word level to facilitate interpretation.

**Sentence-level Aggregation** Following standard sentence-level AA practices, we aggregate token-level citations as the union over all cited documents  $\text{docid}(\cdot)$  across context-sensitive tokens in  $\mathbf{y}$ :

$$\begin{aligned}\text{Mirage}(\mathbf{y}) &= \bigcup_{y_i \in \text{CTI}(\mathbf{y})} \text{docid}(c_j) \quad \forall c_j \in \text{CCI}(y_i) \\ \text{with } s_{\text{cti}} &= m_i \geq m^*, s_{\text{cci}} = a_j^i \geq a^{i*}\end{aligned}$$

In the following sections, we use  $\text{MIRAGE}_{\text{cal}}$  and  $\text{MIRAGE}_{\text{ex}}$  to refer to sentence-level answer attribution using  $m_{\text{cal}}^*$  and  $m_{\text{ex}}^*$  thresholds, respectively.

## 5.4 Agreement with Human Answer Attribution Annotations

We begin our evaluation by comparing MIRAGE predictions to human-produced answer attributions. Importantly, our aim is not to compare several AA approaches to claim optimal faithfulness, but rather to evaluate how our proposed framework fares against existing approaches at the task of producing answer attributions from model internals. We employ the XOR-AttriQA dataset (Muller et al., 2023), which, to our knowledge, is the only open dataset with human annotations over RAG outputs produced by a publicly accessible LM.<sup>5</sup>

We limit our assessment to open-weights LLMs to ensure that MIRAGE answer attribution can faithfully reflect the model’s inner processing towards the natural production of the annotated answer used for evaluation. While these answers could be force-decoded from an open-source model to enable MIRAGE usage, such a procedure would likely impact the validity of AA, as the selected model would not naturally generate the forced answers. Moreover, while cross-linguality is not the focus of our work, XOR-AttriQA allows us to assess the robustness of MIRAGE across several languages and its agreement with human annotations compared to an entailment-based system.

### 5.4.1 Experimental Setup

XOR-AttriQA consists of 500/4720 validation/test tuples, each containing a concise factual query  $\mathbf{q}$ , a set of retrieved documents that we use as context  $\mathbf{c} = \langle \text{doc}_1, \dots, \text{doc}_k \rangle$ , and a single-sentence answer  $\mathbf{y}$  produced by an mT5-base model (Xue et al., 2021) fine-tuned on cross-lingual QA in a RAG setup (CORA; Asai et al. (2021)). Queries

---

<sup>5</sup>E.g., the human-annotated answers in Bohnet et al. (2022) were generated by PALM 540B (Anil et al., 2023), whose internals are inaccessible. See Section A.3.1 for a comparison.

Dataset	BN	FI	JA	RU	TE	Total
Orig.	1407	659	1066	954	634	4720
Match	274	214	232	254	170	1144

**Table 5.2 |** Statistic for test sets of the original XOR-AttriQA and XOR-AttriQA<sub>match</sub>.

and documents span five languages—Bengali (BN), Finnish (FI), Japanese (JA), Russian (RU), and Telugu (TE)—and cross-lingual retrieval is allowed.

Although the RAG generator employs a set of retrieved documents during generation, human annotators were asked to label tuples  $(\mathbf{q}, \text{doc}_i, \mathbf{y})$  to indicate whether the information in  $\text{doc}_i$  supports the generation of  $\mathbf{y}$ .

Notably, MIRAGE requires extracting model internals in the naturalistic setting that leads to the generation of the desired answer, i.e., the one assessed by human annotators. Hence, we perform a selection procedure to identify XOR-AttriQA examples where the answer produced by filling in the concatenated documents  $\mathbf{c}$  in the LM prompt matches the one provided. The resulting subset, which we dub XOR-AttriQA<sub>match</sub>, contains 142/1144 calibration/test examples and is used for our evaluation. Replicating the original answer generation process is challenging since the original ordering of the documents  $\text{doc}_i$  in  $\mathbf{c}$  is unavailable.<sup>6</sup> To maximize the chances of replication, we attempt to restore the original document sequence by randomly shuffling the order of  $\text{doc}_i$ s until LLM can naturally predict the answer  $\mathbf{y}$ . The procedure adopted is described in Algorithm 2. The statistics of the original XOR-AttriQA and XOR-AttriQA<sub>match</sub> are shown in Table 5.2.

#### 5.4.2 Entailment-based Baselines

Muller et al. (2023) use an mT5 XXL model fine-tuned on NLI for performing answer attribution on XOR-AttriQA. Since neither the tuned model nor the tuning data are released, we opt to use TRUE (Honovich et al., 2022), a fine-tuned T5 11B model (Raffel et al., 2020), which was shown to highly overlap with human annotation on English answer attribution tasks (Muller et al., 2023; Gao et al., 2023a). We evaluate TRUE agreement with human annotation in two setups. In NLI<sub>orig</sub>, we evaluate the model directly on all examples, including non-English data. While this leads the English-centric TRUE model out of distribution, it accounts for real-world scenarios with noisy data, and can be used to assess the robustness of the method in less-resourced settings. Instead, in NLI<sub>mt</sub>, all queries and documents are machine translated to English

---

<sup>6</sup>Muller et al. (2023) only provide the split documents without the original ordering.

---

**Algorithm 2** Restore document sequence producing the original annotated answer in XOR-AttriQA

---

```

Require:  $\{Doc_1, \dots, Doc_n\}$ ,  $query$ ,  $answer$ ,  $\mathbb{M}$ 
1: procedure RESTORESEQUENCE( $\{Doc_1, \dots, Doc_n\}$ ,  $query$ ,  $answer$ ,  $\mathbb{M}$ )
2:    $iter = 0$ ,  $found = False$ 
3:   while  $iter < 200$  do
4:      $pred = \mathbb{M}(\{Doc_1, \dots, Doc_n\}, query)$ 
5:     if  $pred == answer$  then
6:        $found = True$ , break
7:     else
8:        $Shuffle(\{Doc_1, \dots, Doc_n\})$ 
9:     end if
10:     $iter += 1$ 
11:   end while
12:   if  $found$  then
13:     return  $\{Doc_1, \dots, Doc_n\}$ 
14:   end if
15: end procedure

```

---

using the Google Translate API. While this simplifies the task by ensuring all TRUE inputs are in English, it can lead to information loss due to imprecise translation.

#### 5.4.3 Results and Analysis

**Mirage agrees with human answer attribution** Table 5.3 presents our results. MIRAGE is found to largely agree with human annotations on XOR-AttriQA<sub>match</sub>, with scores on par or slightly better than those of the ad-hoc NLI<sub>mt</sub> system augmented with automatic translation. Although calibration appears to generally improve MIRAGE’s agreement with human annotators, we note that the uncalibrated MIRAGE<sub>ex</sub> achieves strong performances despite having no access to external modules or tuning data. These findings confirm that the inner workings of LMs can be used to perform answer attribution, yielding performances on par with supervised answer attribution approaches even in the absence of annotations for calibration.

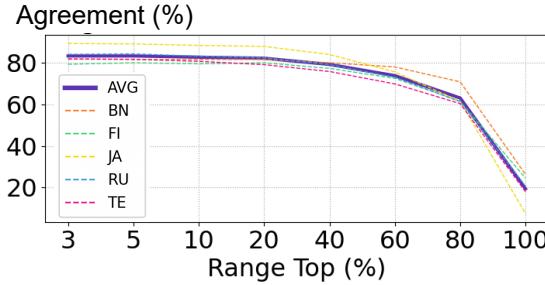
**Mirage is robust across languages and filtering procedures** Table 5.3 shows that NLI<sub>orig</sub> answer attribution performances are largely language-dependent due to the unbalanced multilingual abilities of the TRUE NLI model. This highlights the brittleness of entailment-based approaches in OOD settings, as discussed in Section 5.2.1. Instead,

Method	Extra Req.	CCI Filter	BN	FI	JA	RU	TE	Avg. / Std
NLI <sub>ORIG</sub>	11B NLI model	–	33.8	83.7	86.5	<b>85.8</b>	50.0	68.0 / 21.9
NLI <sub>MT</sub>	11B NLI model + MT	–	82.6	83.7	90.5	81.7	82.5	84.2 / 3.2
MIRAGE <sub>CAL</sub>	142 AA ex.	Top 3 Top 5%	81.7 <b>84.4</b>	87.8 83.0	83.3 <b>91.4</b>	87.0 <b>85.8</b>	86.7 <b>88.9</b>	84.8 / 2.3 86.7 / 3.1
MIRAGE <sub>EX</sub>	–	Top 3 Top 5%	80.2 <b>81.7</b>	78.5 80.1	83.8 <b>89.2</b>	77.2 84.4	75.2 81.8	79.0 / 2.9 83.4 / 3.2

**Table 5.3** | Agreement % of MIRAGE and entailment-based baselines with human AA on XOR-AttriQA<sub>match</sub> using CORA for RAG. **Extra Req.:** data/models needed for AA in addition to the RAG model and the current example. **Filter:**  $s_{cci}$  filtering for saliency scores. **Best overall** and **best uncalibrated scores** are highlighted.

MIRAGE variants perform similarly across all languages by exploiting the internals of the multilingual RAG model. MIRAGE’s performance across languages is comparable to that of NLI<sub>mt</sub>, which requires an extra translation step to operate on English inputs.

We further validate the robustness of the CCI filtering process by testing percentile values between Top 3–100% for the MIRAGE<sub>ex</sub> setting. Figure 5.4 shows that Top % values between 3 and 20% lead to a comparably high agreement with human annotation, suggesting this filtering threshold can be selected without ad-hoc parameter tuning.



**Figure 5.4** | Robustness of MIRAGE<sub>ex</sub> agreement with human annotations across Top-% CCI filtering thresholds.

## 5.5 Answer Attribution for Long-form QA

XOR-AttriQA can only provide limited insights for real-world answer attribution evaluation, as its examples are sourced from Wikipedia articles and its answers are very concise. In this section, we extend our evaluation to ELI5 (Fan et al., 2019), a

challenging long-form QA dataset that was recently employed to evaluate LLM self-citation capabilities (Gao et al., 2023a). Different from XOR-AttriQA, ELI5 answers are expected to contain multiple sentences of variable length, making it especially fitting to assess MIRAGE context-sensitive token identification capabilities before document attribution. Alongside our quantitative assessment of MIRAGE in relation to self-citation baselines, we conduct a qualitative evaluation of the disagreement between the two methods.

### 5.5.1 Experimental Setup

**Dataset** The ELI5 dataset contains open-ended why/how/what queries  $\mathbf{q}$  from the “Explain Like I’m Five” subreddit eliciting long-form multi-sentence answers. For our evaluation, we use the RAG-adapted ELI5 version by Gao et al. (2023a), containing top-5 matching documents  $\mathbf{c} = \langle \text{doc}_1, \dots, \text{doc}_5 \rangle$  retrieved from a filtered version of the Common Crawl (Sphere; Piktus et al. (2021)) for every query. The answer attribution task is performed by generating a multi-sentence answer  $\mathbf{ans} = \langle \mathbf{y}_1, \dots, \mathbf{y}_m \rangle$  with an LLM using  $(\mathbf{q}, \mathbf{c})$  as inputs, and identifying documents in  $\mathbf{c}$  supporting the generation of answer sentence  $\mathbf{y}_i$ ,  $\forall \mathbf{y}_i \in \mathbf{ans}$ .

**Models and Answer Attribution Procedure** We select LLaMA 2 7B Chat (Touvron et al., 2023) and Zephyr  $\beta$  7B (Tunstall et al., 2024) for our experiments since they are high-quality open-source LLMs of manageable size. To enable a fair comparison between the tested attribution methods, we first generate answers with inline citations using the self-citation prompt by Gao et al. (2023b). Then, we remove citation tags and use MIRAGE to attribute the resulting answers to retrieved documents. This process ensures that citation quality is compared over the same set of answers, controlling for the variability that a different prompt could produce. For more robust results, we perform generation three times using different sampling seeds, and report the averaged scores. Since human-annotated data is not available, we only assess the calibration-free MIRAGE<sub>ex</sub>.

**Entailment-based Evaluation** Differently from the XOR-AttriQA dataset used in Section 5.4, ELI5 does not contain human annotations of AA. For this reason, and to ensure consistency with Gao et al. (2023a) self-citation assessment, we adopt the TRUE model as a high-quality approximation of expected annotation behavior. Despite the potential OOD issues of entailment-based AA highlighted in Section 5.4, we expect TRUE to perform well on ELI5 since it closely matches the general/scientific knowledge queries in TRUE’s fine-tuning corpora and contains only English sentences.

Model	Answer Attrib.	Citation ↑		
		Prec.	Rec.	F1
Zephyr $\beta$	Self-citation	41.4	24.3	30.6
	Mirage <sub>ex</sub> Top 3	38.3	46.2	41.9
	Mirage <sub>ex</sub> Top 5%	<b>44.7</b>	<b>46.5</b>	<b>45.6</b>
LLaMA 2	Self-citation	<b>37.9</b>	19.8	26.0
	Mirage <sub>ex</sub> Top 3	21.8	<b>29.6</b>	25.1
	Mirage <sub>ex</sub> Top 5%	26.2	29.1	<b>27.6</b>

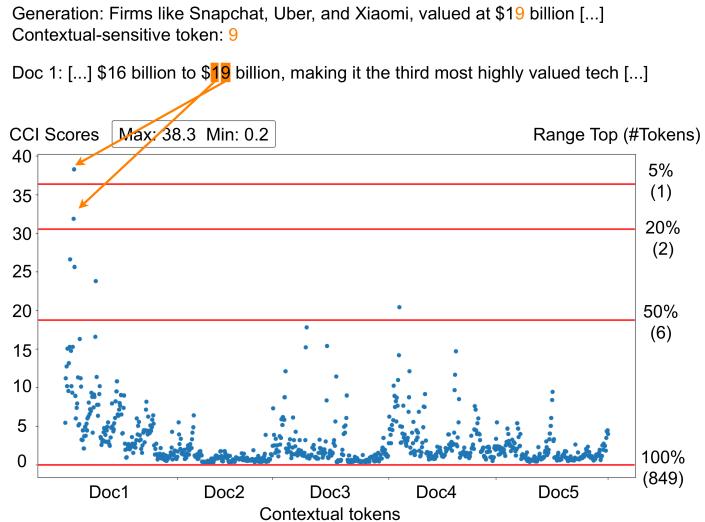
**Table 5.4 |** Answer attribution quality estimated by TRUE for self-citation and MIRAGE on ELI5.

To overcome the multi-hop issue when using single documents for entailment-based answer attribution, we follow the ALCE evaluation (Gao et al., 2023a) to measure citation quality as NLI precision and recall (summarized by F1 scores) over the concatenation of retrieved documents. The ALCE framework for RAG QA evaluation assesses the LLMs’ responses from three viewpoints: citation quality, correctness, and fluency. **Citation quality** evaluates the answer attribution performance with recall and precision scores. The *recall* score calculates if the concatenation of the cited documents entails the generated sentence. The *precision* measures whether each document is cited precisely by verifying if the concatenated text still entails the generation whenever one of the documents is removed. We further calculate F1 scores to summarize the overall performance.

### 5.5.2 Results

Results in Table 5.4 show that MIRAGE provides a significant boost in answer attribution precision and recall for the Zephyr  $\beta$  model. At the same time, it greatly improves citation recall at the expense of precision for LLaMA 2, resulting in an overall higher F1 score for the MIRAGE<sub>ex</sub> Top 5% setting. These results confirm that MIRAGE can produce effective answer attributions in longer and more complex settings while employing no external resources like the self-citation approach.

From the comparison between Top 3 and Top 5% CCI filtering strategies, we note that the latter generally results in better performance. This intuitively supports the idea that an adaptive selection strategy is more suitable for accommodating the wide variability of attribution scores across different examples. Figure 5.5 visualizes the distributions of attribution scores  $a_j^i$  for an answer produced by Zephyr  $\beta$ , showing that most context tokens in retrieved documents receive low attribution scores, with only a handful of



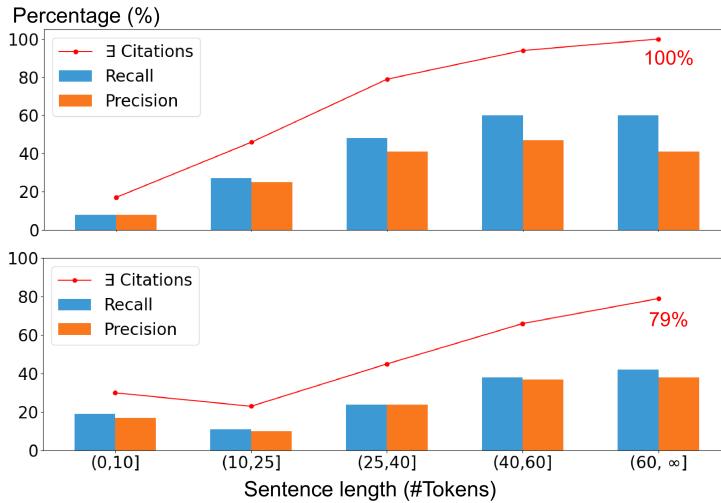
**Figure 5.5 |** Attribution scores over retrieved documents' tokens for the prediction of context-sensitive token ‘9’.

them contributing to the prediction of the context-sensitive token ‘9’ in the generation. This example also provides an intuitive explanation of the robustness of Top-% selection thresholds discussed in Section 5.4.3. Ultimately, the Top 5% threshold is sufficient to select the document containing the direct mention of the generated token.

Since the  $m_{\text{ex}}^*$  threshold used to select context-sensitive tokens by MIRAGE<sub>ex</sub> depends on the mean and standard deviation of generated answer’s scores, we expect that the length of the generated answer might play a role in citation quality. As shown in Figure 5.6, MIRAGE citation quality is indeed lower for shorter answer sentences. However, a similar trend is observed for self-citation, which is outperformed by MIRAGE for all but the shortest length bin ( $\leq 10$  tokens). The proportion of non-attributed sentences (red line) suggests that the lower quality could be a byproduct of the ALCE evaluation protocol, where non-attributed sentences receive zero precision/recall. Future availability of human-annotated RAG datasets may shed more light on this effect.

### 5.5.3 Qualitative Analysis of Disagreements

To better understand MIRAGE’s performance, we examine some ELI5 examples where MIRAGE disagrees with self-citation on Zephyr  $\beta$ ’s generations. Table 5.5 and Table 5.6



**Figure 5.6 |** MIRAGE<sub>ex</sub> (top) and self-citation (bottom) average performance on ELI5 answer sentences binned by length. Red: Percentage of sentences with  $\geq 1$  citations.

illustrate two cases in which the entailment-based TRUE model results agree with either MIRAGE or self-citation.

In Table 5.5, the answer provided by the model is directly supported by Document [1], as also identified by TRUE. However, self-citation fails to cite the related document at the end of the two sentences. By contrast, MIRAGE attributes several spans to Document [1], resulting in the correct answer attribution for both sentences.

While TRUE achieves high consistency with human judgment (e.g., for the example in Table 5.5), NLI-based AA can still prove unreliable in cases of high lexical overlap between the answer and supporting documents. Table 5.6 illustrates one such case, where both self-citation and TRUE attribute the answer to Document [3], whereas MIRAGE does not label any context document as salient for the answer. Here, the answer wrongly states that the bar code can be used to **prevent** the alarm, while Document [3] mentions that the code can be used to **cancel** the alarm after an accidental activation. Thus, despite the high lexical and semantic relatedness, the answer is not supported by Document [3]. The failure of TRUE in this setting highlights the sensitivity of entailment-based systems to surface-level similarity, making them brittle in cases where the model’s context usage is not straightforward. Using another sampling seed for the same query produces the answer “[...] the individual can **cancel** the alarm by providing their password at the keypad”, which MIRAGE correctly attributes to Document [3].

---

INPUT: PROMPT + RETRIEVED DOCS (N=5) + QUERY

---

**Document [1]** [...] Q. What does it mean for books to be Sensitized or Desensitized? A security strip is embedded into each library book. When a book is checked out, it must be "desensitized" so that it will not set off the alarm when it leaves or enters the library. When the book is returned, it is "sensitized" so that the alarm will be set off should someone attempt to take the book from the library without going through the proper borrowing procedure.

**Document [2]** [...]

**Query:** How does a small paper library bar code set off the security alarm when you haven't checked a book out?

---

ANSWER ATTRIBUTION RESULTS

**Self-citation:** [...] When a book is checked out, it is "desensitized" to prevent the alarm from going off.  $\langle \emptyset \rangle$  When the book is returned, it is "sensitized" so the alarm will sound if the item is taken without authorization.  $\langle \emptyset \rangle$  [...]

**Mirage:** [...] When a book is checked<sup>(1)</sup> out, it<sup>(1)</sup> is "desensitized"<sup>(1)</sup> to prevent the alarm from going off. [1] When the book<sup>(1)</sup> is returned, it is "sensitized"<sup>(1)</sup> so the alarm will sound if the item is taken without authorization. [1] [...]

**NLI (True model):** [1] entails both answer sentences.

---

**Table 5.5 |** Example of self-citation failure using Zephyr  $\beta$  on ELI5. NLI and MIRAGE produce the correct citation, while self-citation does not cite any document ( $\langle \emptyset \rangle$ ).

---

INPUT: PROMPT + RETRIEVED DOCS (N=5) + QUERY

---

**Document [2]** [...]

**Document [3]** [...] What will happen if you accidentally set off your security system? The siren will sound and it will be loud, but you should be able to stop the siren by entering your code into your keypad. [...]

**Document [4]** [...]

**Query:** How does a small paper library bar code set off the security alarm when you haven't checked a book out?

---

ANSWER ATTRIBUTION RESULTS

**Self-citation:** [...] False alarms can be prevented by entering the code on the keypad, as documented in [3]. [...]

**Mirage:** [...] False alarms can be prevented by entering the code on the keypad  $\langle \emptyset \rangle$  [...]

**NLI (True model):** [3] entails the answer sentence.

---

**Table 5.6 |** Example showcasing the brittleness of entailment-based AA. MIRAGE correctly finds that the answer cannot be attributed ( $\langle \emptyset \rangle$ ), while NLI and self-citation attribute the lexically similar Document [3].

## 5.6 Limitations

We now highlight some limitations of the MIRAGE method and our experimental evaluation, which should be addressed in future work.

**LLMs Optimized for Self-citation** Our analysis focuses specifically on models that are not explicitly trained to perform self-citation and can provide citations only when prompted to do so. While recent systems incorporate self-citation into their optimization scheme for RAG applications, incorporating model internals into the attribution process will remain a valuable and inexpensive method to ensure faithful answer attributions.

**Brittleness of NLI-based Evaluation** Following Gao et al. (2023a), the evaluation of Section 5.5 employs the NLI-based system TRUE due to the lack of AA-annotated answers produced by open-source LLMs. However, using the predictions of NLI models as AA references is far from ideal, given their brittleness in challenging scenarios and their tendency to exploit shallow heuristics. While the ELI5 dataset is reasonably in-domain for the TRUE model, this factor might still undermine the reliability of some of our quantitative evaluation results. Future work should produce a wider variety of annotated datasets for reproducible answer attribution using open-source LLMs, enabling us to extend our analysis to a broader set of languages and model sizes and ultimately enhance the robustness of our findings.

**Applicability to Other Domains and Models** Our evaluation is conducted on relatively homogeneous QA datasets and does not include language models with >7B parameters. This limits the generalizability of our findings to other domains and larger models. Future work should extend our analysis to a broader range of domains and model sizes to further validate the robustness and applicability of MIRAGE. This said, we expect MIRAGE to be less vulnerable to language and quality shifts compared to existing AA methods that depend on external validators or on the model’s instruction-following abilities.

**Scalability on Longer Context** The computational cost for the simple gradient-based version of MIRAGE we propose is  $2O(F) + |\text{CTI}(y)| \cdot O(B)$ , where  $O(F), O(B)$  are respectively the costs of a forward and a backward pass with the LLM, and  $|\text{CTI}(y)|$  is the number of tokens selected by the CTI step. While CTI effectively limits the expensive backward component in the MIRAGE computation, its cost is bound to increase significantly for larger models and context sizes. When applying MIRAGE to LLMs

with <10B parameters, we note that its cost can be comparable or lower to supervised models like TRUE, requiring several forward passes using a large 11B LLM. Importantly, MIRAGE is a flexible framework that can be implemented using different input attribution methods in the CCI step, including lightweight techniques that require only forward passes, such as Attention Rollout (Abnar and Zuidema, 2020), Value Zeroing (Mohebbi et al., 2023), or ALTI-Logit (Ferrando et al., 2023). Finally, a promising perspective for scaling to larger LLMs could be to assess whether MIRAGE-produced AAs remain accurate when force-decoding the original model’s answer from a different LLM with fewer parameters.

**Parametrization and Choice of Attribution Method** While Section 5.4 highlights the robustness of MIRAGE to various CCI filtering thresholds, the method still requires non-trivial parametrization. In particular, we emphasize that the choice of the attribution method employed to generate attribution scores in the CCI step can significantly impact the faithfulness of the resulting answer attributions. Although we used a relatively simple gradient-based approach, our proposed framework is method-agnostic. We leave the evaluation of modern input attribution techniques, such as the ones mentioned in the previous paragraph, to future work to further improve MIRAGE applicability in real-world settings.

## 5.7 Conclusion

In this chapter, we introduced MIRAGE, a novel approach to enhance the faithfulness of answer attribution in RAG systems. By leveraging model internals, MIRAGE effectively addresses the limitations of previous methods based on prompting or external NLI validators. Our experiments demonstrate that MIRAGE produces outputs that strongly agree with human annotations while being more efficient and controllable than its counterparts. Our qualitative analysis shows that MIRAGE can produce faithful attributions that reflect actual context usage during generation, reducing the risk of false positives motivated by surface-level similarity. Overall, MIRAGE represents a promising first step in exploiting interpretability insights to develop faithful answer attribution methods, paving the way for the usage of LLM-powered question-answering systems in real-world, user-facing applications.

In the next part of this thesis, we will move beyond analysis-driven methods to study how prompting (Chapter 6) and interpretability-based methods (Chapter 7) can be used to effectively condition the machine translation generation process.



## PART II

# Conditioning Generation for Personalized Machine Translation



# CHAPTER 6

---

## Retrieval and Marking for Attribute-Controlled Translation

### Chapter Summary

This chapter shifts the focus from understanding models' context usage to influencing their generation process. As an initial investigation, we introduce Retrieval and Attribute-Marking enhanced Prompting (RAMP), a prompting method that employs semantic similarity retrieval to select in-context examples and attribute marking to identify relevant information, improving generation accuracy with large multilingual language models for few-shot and zero-shot attribute-controlled translation. We experiment on two multilingual datasets for formality and gender-controlled translation, showing that RAMP improves both attribute accuracy and translation quality over standard prompting and MT baselines. The cross-lingual effectiveness of our approach, using in-context examples from various languages, demonstrates the potential for scalable personalization methods when few examples are available.

This chapter is adapted from the paper *RAMP: Retrieval and Attribute-Marking Enhanced Prompting for Attribute-Controlled Translation* (Sarti et al., 2023b).

*Like physical events with their causal and teleological interpretations, every linguistic event had two possible interpretations: as a transmission of information and as the realization of a plan.*

– Ted Chiang, *Stories of Your Life and Others* (2002)

Formality-Controlled Translation (CoCoA-MT)	
Neutral Src (EN)	OK, then please follow me to your table.
Formal Ref (JA)	ではテーブルまで私について来てください。
Informal Ref (JA)	ではテーブルまで私について来て。
Gender-Controlled Translation (MT-GenEval)	
Neutral Src (EN)	After retiring from teaching, Cook became a novelist.
Feminine Ref (NL)	Nadat <u>ze</u> stopte met lesgeven, werd Cook <u>schrijfster</u> .
Masculine Ref (NL)	Nadat <u>hij</u> stopte met lesgeven, werd Cook <u>schrijver</u> .

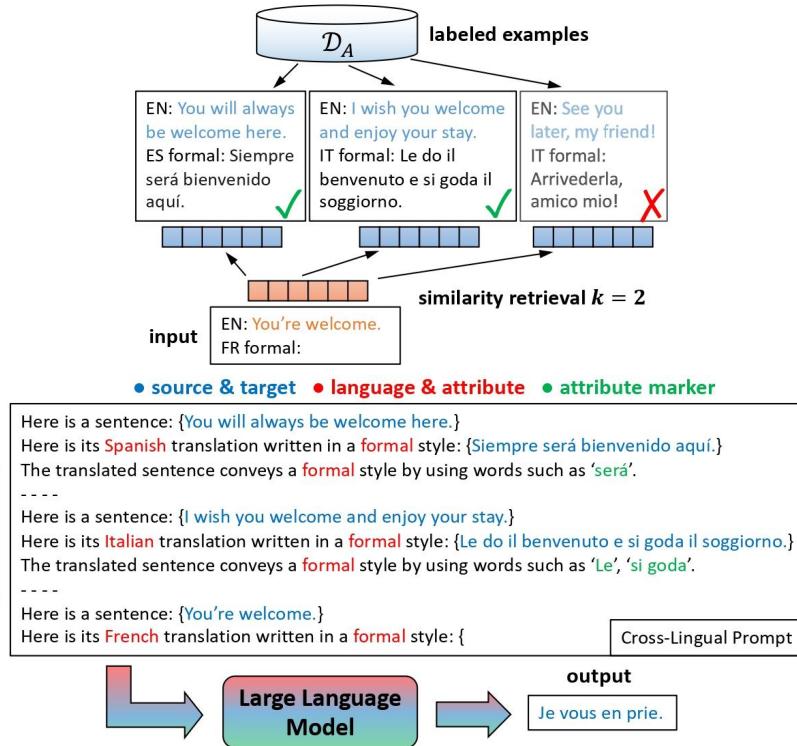
**Table 6.1** | Examples of attribute triplets from CoCoA-MT and MT-GenEval. Attribute markers in the attribute-controlled translations are underlined.

## 6.1 Introduction

*Text style transfer* (TST) is a task that aims to control stylistic attributes of an input text without affecting its semantic content (Jin et al., 2022). Research in TST has focused mainly on English, thanks to the availability of large monolingual English datasets that cover stylistic attributes such as formality and simplicity (Rao and Tetreault, 2018; Zhu et al., 2010). In recent years, however, multilingual and cross-lingual applications of TST have steadily gained popularity (Briakou et al., 2021; Garcia et al., 2021; Krishna et al., 2022). A notable instance of cross-lingual TST is *attribute-controlled translation* (ACT), in which attribute<sup>1</sup> conditioning is performed alongside machine translation to ensure that translations are not only correct but match user-specified preferences, such as formality/honorifics (Sennrich et al., 2016a; Niu et al., 2017; Michel and Neubig, 2018; Niu and Carpuat, 2020; Nadejde et al., 2022; Wang et al., 2023c), gender (Rabinovich et al., 2017; Vanmassenhove et al., 2018; Saunders and Byrne, 2020), and length (Lakew et al., 2019; Schioppa et al., 2021). ACT is crucial for sectors such as customer service and business communication, where stylistic differences can significantly impact user perception (e.g., misgendering customers or speaking to them in an inappropriately informal tone can be perceived as offensive or disconcerting). Table 6.1 shows examples of ACT for formality and gender attributes.

Most prior work on ACT relies on a supervised adaptation component that conditions the generative model on the selective attribute. However, few annotated ACT datasets are available, and they generally cover only a limited set of languages and attributes.

<sup>1</sup>We employ the term *attribute* rather than *style*, since not all the attributes addressed here (e.g., gender) can be considered styles.



**Figure 6.1** | An example of RAMP using 2 in-context examples. **Top:** A sentence similarity model embeds the input sentence, and the top- $k$  most similar labeled examples are retrieved from a pool of training data to build the prompt context. **Bottom:** Labeled cross-lingual examples are used to fill in the English prompt template, which is then provided to the LLM to generate the output.

Thus, enabling few-shot or zero-shot ACT would facilitate applying attribute control to less-resourced attributes and languages.

As a first step into our investigation of conditioning machine translation generation, this chapter introduces a new approach for ACT: **R**etrieval and **A**ttribute-**M**arking enhanced **P**rompting (RAMP). Recent studies have shown that large language models can perform MT out of the box using the prompting paradigm (Brown et al., 2020; Lin et al., 2022; Chowdhery et al., 2023). We build on this, prompting LLMs to perform *attribute-controlled* MT through two innovations: (1) *retrieval of similar examples* and (2) *explicit attribute marking*.

Recent works adopting the prompting paradigm for text style transfer have mainly focused on the generalization capabilities of large English-centric LMs for zero-shot style transfer using previously unseen style descriptions (Suzgun et al., 2022; Reif et al., 2022). However, prior work on other NLP tasks has shown that cross-lingual prompting of multilingual LLMs can be effective (Zhao and Schütze, 2021; Zhou et al., 2023; Huang et al., 2022). As such, we leverage multilingual LLMs and extend their ACT capabilities cross-lingually to languages not covered by the in-context examples, thus enabling zero-shot ACT.

## 6.2 Method

**Attribute-Controlled Translation** ACT takes two inputs, a sentence  $\mathbf{x}$  and a desired target attribute  $a \in A$  (with  $A$  being the space of attributes), and outputs a translation  $\mathbf{y}$  that complies with the specified attribute. It can be formulated as a function  $f : (\mathbf{x}, a) \rightarrow \mathbf{y}$ . In our experiments, we use attribute values provided by the CoCoA-MT formality translation dataset and the MT-GENEVAL gender translation dataset, i.e.,  $A = \{\text{formal, informal}\}$  or  $\{\text{female, male}\}$ .<sup>2</sup>

**Prompting** In the prompting paradigm for decoder-only LLMs, inputs are given as decoding prefixes to the model, usually combined with natural language instructions for output generation. In style-controlled translation, we formulate the prompt for target language  $l$  and attribute  $a$  using the text “*Here is a sentence: { $\mathbf{x}$ } Here is its  $l$  translation written in a  $a$  style.*” to produce the output  $\mathbf{y}$ .<sup>3</sup> In the few-shot setting, we provide a sequence of  $k$  labeled *in-context examples* before the unlabeled input, which can be formulated as a function  $f : \{(\mathbf{x}_1, l_1, a, \mathbf{y}_1), \dots, (\mathbf{x}_{k+1}, l_{k+1}, a)\} \rightarrow \mathbf{y}_{k+1}$ .

### 6.2.1 Our Approach: RAMP

RAMP builds on the success of the prompting paradigm on few-shot generation tasks such as monolingual text style transfer (Reif et al., 2022) and MT (Garcia and Firat, 2022; Agrawal et al., 2023) by creating more informative prompts through *similarity retrieval* and *attribute marking*. See Figure 6.1 for an illustration of RAMP.

**Similarity Retrieval** In standard prompting, in-context examples are sampled randomly from the pool of labeled examples  $\mathcal{D}_A$ . In RAMP, we select examples based on

---

<sup>2</sup>See Section 6.4 for ethical considerations.

<sup>3</sup>We adopt prompt templates similar to the one used by Reif et al. (2022), and we write the prompt template in English. Complete templates are provided in Section B.1.1.

their similarity with the input text. We first embed both the input text and the source texts of  $\mathcal{D}_A$  using all-MiniLM-L6-v2 (Wang et al., 2020). Then, the top- $k$  most similar examples are retrieved for the input text based on cosine similarity. These are then used in a descending order based on their cosine similarity as in-context examples in the inference prompt. As demonstrated in Figure 6.1, the in-context example “You will always be welcome here.” has the highest similarity to the test example “You’re welcome.”, so it is prompted first.

**Attribute Marking** In standard prompting, in-context examples are provided without explicit information on why they satisfy the prompting objective. Inspired by recent studies that have shown that decomposition of complex tasks can improve prompting quality (Nye et al., 2022; Wei et al., 2022), we include for every in-context example an additional sentence directly after the target sentence that specifies which text spans convey the desired attribute (e.g., *“The translated sentence conveys a formal style by using words such as ‘Vous.’”*). In our experiments, we use the gold attribute spans included in the CoCoA-MT and MT-GenEval datasets. In Section 6.5 we suggest possibilities for automatically deriving attribute spans when gold training labels are not available.

#### 6.2.2 Cross-Lingual Prompting

The similarity retrieval component of RAMP requires a large pool  $\mathcal{D}_A$  from which to find appropriate in-context examples for prompting. Low-resource attributes or language pairs may have insufficient or no annotated data from which to retrieve such examples. To mitigate this issue, we introduce *cross-lingual prompting*, in which the target side of the in-context examples differs from the desired target language of the translation task. As demonstrated in Figure 6.1, we investigate whether the system can leverage examples in one language (e.g., attribute indicators in Spanish) to produce the same attribute in another (e.g., French). Two main features of our RAMP model allow us to perform cross-lingual prompting: (1) the use of multilingual LLMs, and (2) the example retrieval step, which is done on the source language only.

### 6.3 Experiments

In this section, we describe the datasets, LLMs, and baselines used in our experiments, as well as the evaluation metrics. We then present the results of RAMP in both same-language and cross-lingual prompting settings.

AR	Arabic	DE	German	EN	English
ES	Spanish	FR	French	HI	Hindi
IT	Italian	JA	Japanese	NL	Dutch
RU	Russian				
	AR	ES	FR	HI	PT
CoCoA-MT		✓	✓	✓	✓
MT-GenEval	✓	✓	✓	✓	✓
XGLM	✓	✓	✓	✓	✓
BLOOM	✓	✓	✓	✓	✓

**Table 6.2 | Top:** Language codes used in the RAMP experiments. **Bottom:** Target languages in the test sets and languages **seen** by LLMs in pre-training. We report results on languages seen by both LLMs.

### 6.3.1 Datasets

We experiment on two multilingual ACT datasets:

- **CoCoA-MT** (Nadejde et al., 2022) covers formality-controlled translation in the conversation domain. Source sentences are underspecified for formality, and references require formality markings (formal or informal).
- **MT-GENEVAL** (Currey et al., 2022) covers gendered translation in the Wikipedia domain. We use the *contextual* subset, in which sentences are gender ambiguous in the source while the reference requires gender marking. We do not use the disambiguating sentences; instead, we explicitly control the target gender.

Both datasets have gold annotations for attribute-marked target spans, and both cover translation from English into multiple diverse target languages. We list their target languages in Table 6.2.

### 6.3.2 Large Language Models

We select three massively multilingual decoder-only LLMs for the prompting experiments:

- XGLM (Lin et al., 2022) is a 7.5B-parameter model trained on a balanced corpus containing 30 languages. It was shown to outperform much larger models such as GPT-3 on tasks related to machine translation and cross-lingual language understanding. We select it due to its broad linguistic coverage and its manageable size.

Dataset	Attribute	# Train	# Test	Acc.
CoCoA-MT	Formality	7,600	1,596	0.990
MT-GenEval	Gender	4,900	9,854	0.970

**Table 6.3 |** Dataset statistics. We report # of triplets in the train/test split aggregated across all languages and the classification accuracy on the test split of the classifiers.

- BLOOM (BigScience Workshop et al., 2022) is a model available in multiple sizes, trained on a curated corpus spanning 46 natural languages (and 13 programming languages). However, many of the test languages are not part of its pre-training corpus (see Table 6.2). We evaluate two variants of the model (7.1B and 175B parameters) to assess how it is affected by a massive scaling in model parameters. The larger variant has a parameter count comparable to that of GPT-3, making it the largest publicly available multilingual LLM at present.
- GPT-NEOX (Black et al., 2022) is a 20B-parameter model trained on The Pile (Gao et al., 2021), a large English-centric corpus covering a broad range of domains. While the model was primarily trained on English data and is therefore not intended for multilingual usage, it exhibits interesting generalization performance for many of our target languages.

The selected models span three orders of magnitude in terms of number of parameters and differ in the languages that they cover (see Table 6.2).

### 6.3.3 Baseline

Attribute tagging is a standard method for ACT, so we include a baseline following the approach and configuration used by Nadejde et al. (2022), i.e. an encoder-decoder transformer MT model (Vaswani et al., 2017) pre-trained on public parallel data and further finetuned on contrastive training pairs with attribute tags (from either CoCoA-MT or MT-GENEVAL) such as `<formal>`, `<informal>`, `<mASCULINE>` and `<fEMININE>`. We refer to these models as **adapted MT** in our evaluation.

### 6.3.4 Evaluation Metrics

We measure translation quality with BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020). For attribute accuracy, we use the lexical matching metrics provided with CoCoA-MT and MT-GENEVAL (**Lexical-Accuracy**) and sentence encoders trained on contrastive examples (**Sentential-Accuracy**). For the latter, we train multilingual

classifiers on top of the mDeBERTa-v3 encoder (He et al., 2023). High-performance pre-trained classifiers have been shown to produce attribute accuracy estimates closer to human judgments for style transfer (Lai et al., 2022). Table 6.3 presents the accuracy of the classification models on the test sets of their respective datasets, averaged across all languages.

We use the original train/test split provided by the CoCoA-MT dataset. Each split contains *telephony* and *topical\_chat* domains. We use the *topical\_chat* domain in our experiments. MT-GENEVAL contains a dev and test split, and we use the dev split as training data for the classification model and prompting experiments.

We finetune mDEBERTA-v3-BASE model<sup>4</sup> on the contrastive examples in the respective training sets to obtain the attribute classifiers. We fine-tune the classifier for two epochs with a batch size of 8, a learning rate of 2e-5, 500 warm-up steps, a max sequence length of 256, and save checkpoints every 500 steps. We do not do hyperparameter tuning, and thus, a validation set is not used.

Unlike lexical accuracy, the multilingual attribute classifier does not penalize text generated in incorrect languages. Thus, in cross-lingual prompting experiments, we include a step of language detection<sup>5</sup> so that generated sentences not in the requested target language are considered incorrect.

### 6.3.5 Same-Language Prompting

We first evaluate the effectiveness of RAMP for formality- and gender-controlled translation where the language pair used for in-context examples is the same as the one used in the prompt candidate (e.g., English→Spanish formality-controlled translation using English→Spanish in-context examples).

We begin by conducting a preliminary evaluation of 3 LLMs across different ranges of in-context examples to reduce the number of experimental settings for our main assessment. We perform formality-controlled translation using CoCoA-MT, and evaluate LLMs by varying the number of in-context examples (i.e., 4-8-16-32, selected based on the feasible context length<sup>6</sup>). Figure 6.2 presents results averaged across all four languages seen by BLOOM during its pre-training.

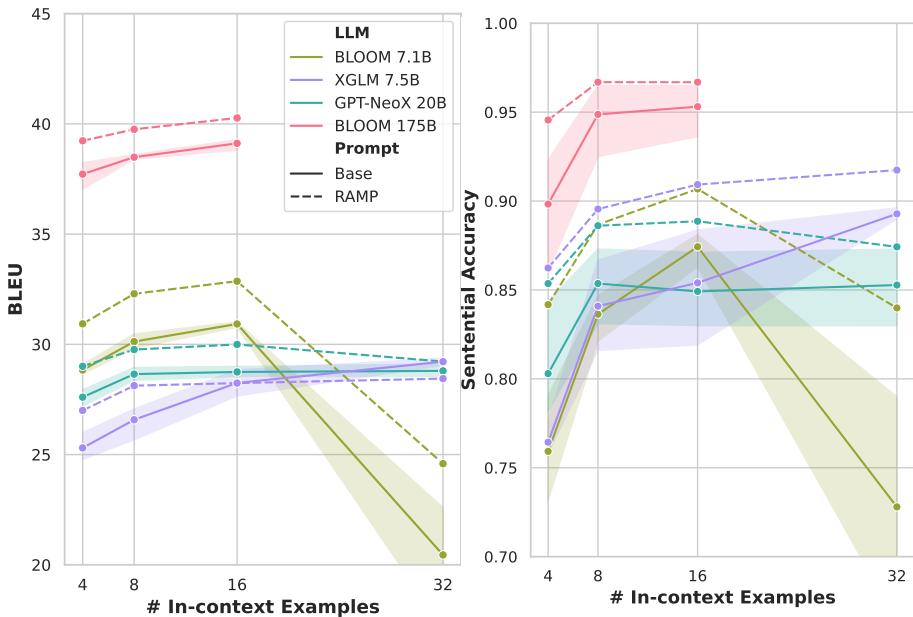
We observe that RAMP generally outperforms base prompting (i.e., random in-context

---

<sup>4</sup>[microsoft/mdeberta-v3-base](https://microsoft/mdeberta-v3-base)

<sup>5</sup><https://pypi.org/project/langdetect/>

<sup>6</sup>BLOOM 175B encountered out-of-memory errors with 32 in-context examples on 8 A100 40GB GPUs.



**Figure 6.2** | BLEU and sentential formality accuracy of prompt outputs on CoCoA-MT test set for different amounts of in-context examples in the base and RAMP settings. Confidence intervals are obtained for the base setting by sampling in-context examples using 3 seeds. Detailed scores are included in Table B.1.

examples and no attribute marking) across most LLMs and example settings for both BLEU and formality accuracy. Moreover, BLEU and formality accuracy improve with increased model size and with the number of examples, until this number reaches 16. Based on these results, we move forward with the main evaluation using XGLM 7.5B and BLOOM 175B models and 16 in-context examples for both datasets.

Table 6.4 presents our main results alongside the adapted MT baseline. The base model uses in-context examples that are randomly sampled from the pool of labeled examples. We also include an ablation that adds only attribute marking on top of base prompting, without similarity retrieval (**+mark**).

We observe that in the **+mark** setting, simple attribute marking consistently improves attribute accuracy of the generated text, but leads to degradation of COMET on CoCoA-MT. The complete RAMP with similarity retrieval not only compensates for the COMET degradation but also improves quality and attribute metrics across the board, especially for the high-capacity BLOOM 175B model.

			CoCoA-MT				MT-GenEval			
			BLEU	COMET	L-Acc	S-Acc	BLEU	COMET	L-Acc	S-Acc
Same Lang.	XGLM	base	28.6	<b>0.463</b>	0.835	0.846	23.7	0.445	0.790	0.727
		+mark	28.7	0.423	0.920	0.902	23.7	0.444	0.789	0.732
		RAMP	<b>30.0</b>	0.451	<b>0.938</b>	<b>0.923</b>	<b>24.8</b>	<b>0.473</b>	<b>0.836</b>	<b>0.820</b>
	BLOOM	base	39.9	0.691	0.930	0.940	33.3	0.679	0.748	0.704
		+mark	40.3	0.688	0.970	<b>0.970</b>	33.1	0.674	0.759	0.725
		RAMP	<b>41.9</b>	<b>0.711</b>	<b>0.973</b>	<b>0.970</b>	<b>34.3</b>	<b>0.699</b>	<b>0.817</b>	<b>0.818</b>
	Adap. MT		38.5	0.454	0.691	0.693	39.6	0.750	0.842	0.864
Cross Lang.	BLOOM	base	32.1	0.644	0.567	0.596	28.5	0.469	0.777	0.633
		RAMP	31.8	0.646	<b>0.625</b>	<b>0.622</b>	<b>29.4</b>	<b>0.502</b>	<b>0.788</b>	<b>0.673</b>

**Table 6.4 |** BLEU, COMET, Lexical- and Sentential-Accuracy of selected LLMs using 16 same-language in-context examples on two tasks, alongside adapted MT models. Scores are aggregated across *seen* languages (w.r.t. BLOOM pre-training) and both attributes for each task. (Full results in Table B.2–Table B.5.)

Adapted MT outperforms BLOOM 175B on MT-GENEVAL in all metrics, but underperforms it on CoCoA-MT. This suggests that it is challenging to conduct a fine-grained comparison between LLMs and standard MT systems, as they may have different domain coverage. BLOOM 175B consistently outperforms XGLM 7.5B in both generic translation quality and attribute control accuracy, so we focus on BLOOM 175B for our cross-lingual prompting analysis.

### 6.3.6 Cross-Lingual Prompting

We have demonstrated the effectiveness of selecting similar same-language examples to build the prompt, echoing related work (Liu et al., 2022; Agrawal et al., 2023). In this section, we evaluate the cross-lingual prompting option, i.e., retrieving in-context examples from other target languages besides the desired language of translation. We test this zero-shot setting using the leave-one-out strategy, i.e. we retrieve in-context examples from every language except the desired language of translation. We ensure that we retrieve an equal number of examples from all languages: the number of examples retrieved from each language is the total desired number of in-context examples divided by the number of training languages. In CoCoA-MT, we retrieve 14 in-context examples from 7 languages. In MT-GENEVAL, we retrieve 8 in-context examples from 8 languages.<sup>7</sup> Finally, results are averaged across tested language pairs. Languages that

<sup>7</sup>We reduced the number of in-context examples in this setting to avoid out-of-memory errors with BLOOM 175B.

are not seen during the LLM pre-training are included among in-context examples, but not as the target language of the translation task.

Table 6.4 (bottom) presents our results using BLOOM 175B. On both test sets, compared to the baseline, we observe improved attribute accuracy and comparable or better generic translation quality when using RAMP with cross-lingual prompting.

We observe translation quality degradation with RAMP on some target languages of CoCoA-MT, such as Spanish. Manual analysis shows that repeatedly inaccurate retrieval results could lead to hallucinations.<sup>8</sup> For example, RAMP retrieves multiple sentences containing “million” for the input `If you got it why not? He is worth over 20 billion dollars after all.` This results in mistranslation of *billion* to *million (millionario)*: `Si lo tienes, ¿por qué no? Es millonario después de todo.` We give detailed examples in Section B.1.3. This is a known issue with retrieval-based prompting (Liu et al., 2022; Agrawal et al., 2023), which can be mitigated by using more diverse in-context examples or a larger pool of training data for retrieval.

## 6.4 Limitations

We delineate some limitations of our approach and discuss future work directions.

**Example Availability and Prompt Sensitivity** The proposed formulation of the RAMP method relies on gold annotations for attribute marking, which are not always available depending on the dataset. However, RAMP could be easily extended to unsupervised settings through LLM input attribution methods, such as those we presented in the previous part of this thesis. This approach builds upon recent techniques in unsupervised language generation metrics (Fomicheva et al., 2021, 2022a; Leiter et al., 2024). Apart from the choice of in-context examples, prompting is also sensitive to their ordering (Lu et al., 2022) and the design of the template (Jiang et al., 2020). We refrain from tuning example orders and templates to avoid introducing too many variables, but we acknowledge that this could lead to suboptimal results.

**Unseen Languages, Computational Resources and Diversity** Multilingual LLMs perform competitively on machine translation for languages seen during their pre-training. However, we noticed that BLOOM 175B produces better English→Italian translations than XGLM 7.5B even though Italian is not listed among BLOOM’s training

---

<sup>8</sup>Vilar et al. (2023) also observe hallucinations when the retrieved examples have bad translations (i.e., non-parallel sentences).

languages. This could be due to typological similarity between Italian and the Romance languages included in BLOOM training. Multilingual LLMs such as BLOOM also require significantly more GPU resources for inference than standard bilingual MT systems do, making them less practical for production deployment. Finally, the MT-GENEVAL test set is limited in providing only two gender labels (`female` and `male`) as minimal pairs, while neutral rewriting is not represented.

## 6.5 Conclusion

As a first step in our exploration of conditioning machine translation generation, we introduced the RAMP in-context learning for better conditioning performance through the use of attribute annotations and similar retrieved examples. We demonstrated its effectiveness with multilingual LLMs for both formality-controlled and gender-controlled translation, showing that it improves attribute accuracy and translation quality over standard prompting and adapted MT baselines, including in cross-lingual settings using relevant in-context examples from other languages. In the next chapter, we expand our investigation to steering methods that intervene directly in the inner processing of LLMs and study their effectiveness for personalization in the challenging domain of literary machine translation.

# CHAPTER 7

---

## Steering Language Models for Personalized Machine Translation

### Chapter Summary

This chapter expands our evaluation of machine translation conditioning approaches by assessing the effectiveness of efficient inference-time interventions on model internals for personalizing large language models' outputs. Focusing on the challenging domain of literary translation, we explore prompting strategies and inference-time interventions using sparse autoencoders to steer model generations toward personalized translator styles. We propose a contrastive framework that exploits interpretable latent concepts from SAEs to identify salient personalization properties, and demonstrate that its strong personalization accuracy, comparable to and at times better than few-shot prompting, does not come at the cost of translation quality. Our analyses further reveal that successful SAE steering and multi-shot prompting impact similar model layers, suggesting similar mechanisms at play.

This chapter is adapted from the paper *Steering Large Language Models for Machine Translation Personalization* (Scalena, Sarti, et al., 2025).

*I don't speak, I operate a machine called language. It creaks and groans, but is mine own.*

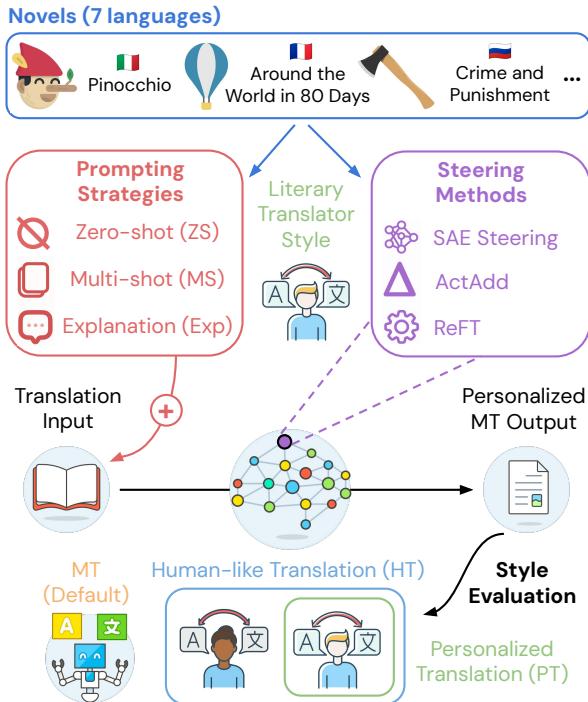
– Frank Herbert, *Dune Messiah* (1969)

## 7.1 Introduction

When we read a translated book, we do not simply read the story in a new language; we also experience the translator’s personal voice through their stylistic choices. Past efforts in the automatic translation of literary works have historically been constrained by the limited capabilities and flexibility of machine translation systems. The recent popularization of MT systems based on large language models has significantly improved their capacity to handle the long contexts typical of literary translations, but mimicking the creative and rich language that characterizes the translators’ own style remains an open issue. In this context, several works have explored the use of prompting and tuning-based strategies to ensure that translations are stylistically appropriate (Michel and Neubig, 2018; Wang et al., 2021). However, their influence on model internal representations is rarely explored, making their impact less controllable and often unpredictable. Building upon the prompting techniques demonstrated in Chapter 6, this chapter tackles the more complex challenge of personalizing machine translation to match individual translator styles. While RAMP focused on explicit attributes like formality and gender, literary translation requires capturing the subtle, implicit stylistic preferences that characterize individual translators’ voices. For this purpose, we compare prompting approaches with *steering* methods proposed in interpretability literature. These techniques can be used to surgically intervene on LLMs’ intermediate representation to generate personalized translations when few examples are available, using the PAR3 dataset (Thai et al., 2022) with multiple human translations for novels translated into English from 7 typologically diverse languages.

We begin with preliminary assessments by verifying whether translators’ styles are discernible by automatic systems, finding that trained classifiers can distinguish writing styles with high accuracy, while the task is notoriously challenging for human annotators (Youyou et al., 2015; Flekova et al., 2016). We also find a simple prompting setting with in-context personalization examples to improve the style accuracy of LLM translation, suggesting personalized translation styles are reproducible. We connect the conditioning induced by prompting to the inner workings of the model, identifying activations with high discriminative capacity for style differences in intermediate model layers. We then propose a contrastive steering approach based on sparse autoencoders (SAEs, Huben et al. (2024)) to condition model generations by upweighting sparse, interpretable latents at inference time. We validate the effectiveness of our method across three LLMs of various sizes on PAR3 novels, comparing our results with established prompting and steering methods.

Our results show that contrastive SAE steering is a promising approach for MT person-



**Figure 7.1** | We compare prompt-based approaches with steering techniques intervening on model internals for personalizing MT outputs in literary machine translation, employing MT quality metrics and style classifiers to disentangle the effect of steering on outputs fluency and personalization adequacy.

alization, resulting in translations that not only align more closely with general human translation features but also with the desired personalized style compared to other methods. Importantly, these results are achieved with no degradation in translation quality, according to established MT quality metrics. We conclude by comparing the impact of our method on model representations with the outcome of multi-shot prompting, finding that probes trained on prompt-conditioned activations can predict the effectiveness of SAE steering with high precision. These results confirm that tested prompting and steering techniques converge to similar solutions for conditioning model behavior, enabling future investigations into the mechanistic impact of prompting through the study of learned SAE latents and other interpretable components.

## 7.2 Related Work

**Machine Translation of Literary Texts** The literary domain has historically been challenging for automatic MT systems due to their limited ability in handling rich linguistic and cultural contexts (Matusov, 2019) and their propensity to produce overly literal outputs (Guerberof-Arenas and Toral, 2022). Automatic literary translation has a long history dating back to pre-neural MT approaches (Voigt and Jurafsky, 2012; Toral and Way, 2015, 2018; Moorkens et al., 2018) with two recent dedicated evaluation campaigns (Wang et al., 2023b, 2024a). The advent of LLMs has brought new opportunities in processing longer contexts for document-level translation (Wang et al., 2023a; Briakou et al., 2024; Wu et al., 2025), but critical errors requiring human translators’ intervention nonetheless persist (Karpinska and Iyyer, 2023). Here, we use the PAR3 dataset (Thai et al., 2022) containing multiple human translations of novels to evaluate MT personalization in the literary domain.

**Personalization for Machine Translation** Advances in MT quality recently led to a growing interest in personalization approaches to ensure a consistent format and appropriate stylistic choices in model generations (Rabinovich et al., 2017; Lin et al., 2021). Previous approaches for controlling attributes such as formality (Sennrich et al., 2016a; Niu et al., 2017; Nadejde et al., 2022) or gender (Vanmassenhove et al., 2018; Saunders and Byrne, 2020) typically required tuning existing models on pre-defined properties of interest, with few works attempting a real data-driven adaptation from unlabeled demonstrations (Michel and Neubig, 2018; Wang et al., 2021; Zhang et al., 2022). More recently, several studies employed prompting (Garcia and Firat, 2022; Sarti et al., 2023b) or preference optimization from post-editing behavior (Lee et al., 2023a; Berger et al., 2024) to render MT personalization more effective and data-efficient. We complement established prompt methodologies with steering approaches to personalize MT outputs using a few user-provided examples.

## 7.3 Preliminaries

Before testing the effectiveness of personalization strategies, we validate some key assumptions: **i)** Whether the personalized translation style is *discernible*, i.e., if it is possible to tell apart human- and machine-generated translations; **ii)** Whether different translation styles are automatically *reproducible*, i.e., if LLMs can mimic a specific translator’s style when provided with some examples; and **iii)** Whether style distinctions are reflected in the model’s internal representations, to motivate the interest in steering approaches for personalization.

ZH → EN	道人道：「既如此，便你去。」
H <sub>1</sub>	"Such being the case," the Taoist acquiesced, "I am ready to follow you, whenever you please to go."
H <sub>2</sub>	"Very good, I will go with you then," said the Taoist.
Gemma 2 2B	"If that's the case, then go ahead" said the Taoist.
Gemma 2 9B	"If so, then I will go with you" the Taoist said
Llama 3.1 8B	The Taoist said: "If you insist on going, then go ahead."
IT → EN	Sarà l'effetto dell'acqua del mare. Il mare ne fa di questi scherzi
H <sub>1</sub>	"The salt water must have done it. The sea plays funny tricks."
H <sub>2</sub>	"It must have been the effect of sea-water. The sea makes extraordinary changes."
Gemma 2 2B	"It will be the effect of the sea water. The sea makes of these jokes."
Gemma 2 9B	It will be the effect of the sea water. The sea plays these tricks.
Llama 3.1 8B	It will be the effect of the sea water. The sea does things like this.

**Table 7.1 | ZH→EN and IT→EN examples for PAR3 segments translated by humans (H<sub>1</sub>, H<sub>2</sub>) and LLMs with zero-shot prompting (MT). More examples in Table B.8, Table B.7.**

We use the PAR3 dataset by Thai et al. (2022), which contains multiple non-English novels, as a benchmark to evaluate personalization. Novels are segmented into paragraphs with translations into English by two professional literary translators. To ensure a diverse and representative evaluation, we select novels spanning a variety of linguistic families and cultural backgrounds. Our dataset includes Romance languages such as Italian (*Pinocchio*) and French (*Around the World in Eighty Days*), as well as Germanic languages like Dutch (*The Diary of a Young Girl*) and German (*Beware of Pity*). To evaluate our setup on non-Latin scripts and distinct linguistic structures, we also include Russian (*Crime and Punishment*), Japanese (*No Longer Human*), and Chinese (*Dream of the Red Chamber*). Table 7.2 summarizes the number of paragraphs employed in the evaluation of each language.

Examples for a subset of languages are shown in Table 7.1. We name the two available human translations H<sub>1</sub> and H<sub>2</sub>, and compare them with MT outputs produced by LLMs, which we denote as MT<sub>model</sub>. We use three LLMs, namely Llama 3.1 8B Instruct (Team, 2024b) and Gemma 2 (Team, 2024a) in its 2B and 9B instruction-tuned variants. Our model selection is motivated by our steering requirements, discussed in Section 7.5.

Lang	Novel name	Train	Val	Test
Italian (IT)	Pinocchio	745	82	107
French (FR)	Around the World in Eighty Days	829	92	120
Dutch (NL)	The Diary of a Young Girl	769	85	110
German (DE)	Beware of Pity	606	67	96
Russian (RU)	Crime and Punishment	1517	168	224
Japanese (JA)	No Longer Human	652	40	81
Chinese (ZH)	Dream of the Red Chamber	694	76	92

**Table 7.2 |** Languages and novels used in our evaluation. Numbers corresponds to a single paragraph in the dataset. In training, each paragraph is associated with three distinct translations:  $H_1$ ,  $H_2$ , and  $MT$ . 20 additional in-context examples are left aside for each language for prompting and steering methods.

### 7.3.1 Are Personalized Translations Discernible?

Following prior work on personalization (Wang et al., 2024c; Liu et al., 2023b), we train a series of classifiers based on multilingual XLM transformer encoders (Conneau et al., 2020) to distinguish between  $H_1$ ,  $H_2$ , and  $MT$  translations. If those systems can reliably separate these three classes, it suggests the presence of reasonably distinct stylistic signals differentiating them. In particular, the ability to distinguish between  $H_1$  and  $H_2$  would denote not only the possibility to discern a human-like style from human-made and automatic translations, but also a *personalized style* from different human translators.

We train a classifier for each language and each model in our evaluation suite. All classifiers are fine-tuned from the `xlm-roberta-large` model<sup>1</sup>, using a linear classification head. Training is conducted for 6 epochs with a learning rate of 2e-5 and a batch size of 32, selecting the best model checkpoint based on validation accuracy. Training data only includes generations from models and the translator without any source text. It is also perfectly balanced, as each paragraph provides one instance for all three labels:  $H_1$ ,  $H_2$ , and  $MT$ . The total size of the training set varies depending on the number of paragraphs in the chosen novel. On average, we obtain approximately 830 instances, resulting in a total of around 2,490 labeled examples for training (see Table 7.2). Validation and test sets are strictly held out and never seen during training. Additionally, they do not include the small 20-example subsets used for prompting or steering. Results in Table 7.3 indicate that translation styles are discernible with high accuracy. On average, across all models and languages, the classifiers achieve an accuracy ranging from 77% (Japanese) to 99% (Chinese), with an overall average of 86%. These results suggest

<sup>1</sup> FacebookAI/xlm-roberta-large

Lang.	Gemma 2 2B	Gemma 2 9B	Llama 3.1 8B
DE	0.89	0.90	0.84
RU	0.92	0.90	0.91
ZH	0.99	0.98	0.98
IT	0.78	0.85	0.80
NL	0.79	0.78	0.82
FR	0.88	0.87	0.90
JA	0.76	0.79	0.76

**Table 7.3 |** Accuracy of model- and language-specific 3-way ( $H_1$ ,  $H_2$ ,  $MT$ ) classifiers on balanced held-out sets for every language. Random baseline: 0.33.

that personalization information is abundant in the literary setting and can plausibly be exploited for modeling. These findings corroborate previous results showing the high learnability of this task by machines while remaining intrinsically difficult for human annotators (Youyou et al., 2015; Flekova et al., 2016; Wang et al., 2024b).<sup>2</sup>

### 7.3.2 Can LLMs Reproduce Human Translation Styles?

To confirm whether MT personalization can be achieved, we test the LLM’s ability to mimic the stylistic choices of a particular translator in a multi-shot (MS) prompting setup. For each translator available across tested novels, we provide the model with 20 in-context examples selected from the original pool of translated paragraphs by that translator, asking it to generate a consistent translation. We compare MS results with the default zero-shot (ZS) prompting, which uses no examples from the translator, to quantify the effect of in-context examples. Table 7.4 presents results for *personalization accuracy*, automatically evaluated using our high-scoring classifiers from the previous section; and *translation quality*, estimated via the widely used COMET MT metric (Rei et al., 2020). The proportion of outputs categorized as matching the translator’s style is increased two- to four-fold following MS prompting, suggesting that LLMs can employ implicit clues in small sets of user examples to produce personalized translations. Stable scores for COMET also confirm that translation quality is maintained during style adaptations.

---

<sup>2</sup>Two human annotators were asked to label 100 translated paragraphs from the novel Pinocchio (IT→EN) as either human or MT, resulting in an accuracy of ~ 60%.

	Gemma 2 2B		Gemma 2 9B		Llama 3.1 8B	
ZS	0.10	0.69	0.08	0.71	0.08	0.70
MS	<b>0.24</b>	0.69	<b>0.31</b>	0.73	<b>0.32</b>	0.73

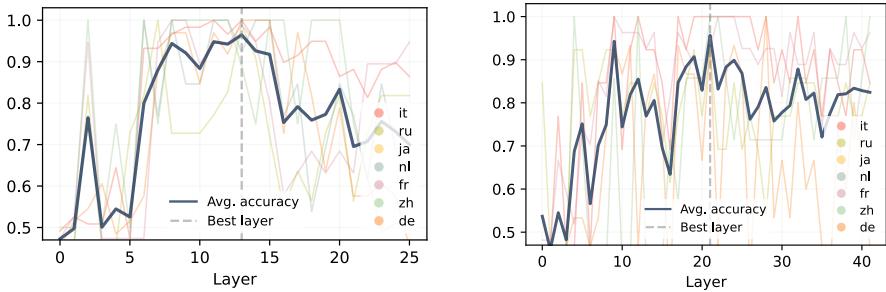
**Table 7.4 |** Classifier-based personalization accuracy () and Comet-based translation quality () for zero-shot (ZS) and multi-shot (MS) prompting with 20 in-context examples averaged across all translators and languages.

### 7.3.3 Finding Personalization Information in LLM Representations

In light of these results, we set out to test how the model encodes information reflecting a stylistic shift when style-appropriate examples are provided. To this purpose, we train *linear probes* (Belinkov, 2022) using model activations as input features to predict the style label (**MT**,  $H_1$ , or  $H_2$ ) that the style classifier (from Section 7.3.1) would assign to the eventual translation, based purely on the prompt’s internal representation. Probing accuracy is measured by testing the model’s ability to predict the classified outcome before generation, using only the prompt representation formed by the model. Given a test set of human-translated paragraphs, we train our probes on a set of examples using an MS prompt with 20 in-context examples. The set is balanced between prompts that showcase personalization with gold in-context examples from a human translator and non-personalized prompts with MT-generated examples previously produced by the same tested model in a ZS setup. Test examples are selected from the respective novels to ensure for the classifier prediction shifts from **MT** in the ZS setting to the style of in-context examples when MS is used, signaling a causal influence of demonstrations on output personalization.<sup>3</sup> This balanced setup prevents the leakage of task information, such as the number of in-context examples, to learned probes, ensuring that stylistic differences between human- and machine-generated in-context examples are the sole factor determining differences in model activations. We focus specifically on Gemma models, extracting activations after the attention block at each model layer for the last token of the prompt, which was previously shown to encode key task-relevant information (Hendel et al., 2023; Todd et al., 2024; Scalena et al., 2024). Figure 7.2 reports probe accuracies across all layers of Gemma 2 2B and 9B. We find a peak in probe accuracy of  $\sim 95\%$  around intermediate model layers, suggesting that these layers encode stylistic information with near-perfect precision.<sup>4</sup> These results confirm

<sup>3</sup>Examples are resampled for every test paragraph to prevent the probe from overfitting on spurious prompt features.

<sup>4</sup>We find probes for layers 13 and 21 to perform best for the 2B and 9B models, respectively.



**Figure 7.2 |** Probing classifier performance on the human translation detection task across Gemma 2 2B (left) and 9B (right) layers. Activations in intermediate layers are found to capture translation style information with high precision.

that personalization is discernible from LLMs’ internal representation, motivating our experiments towards the design of inference-time interventions to steer models towards personalized MT outputs.

## 7.4 Methods

We begin by introducing the prompting and steering methods that we use as baselines and outline our own proposed SAE-based steering approach for personalized translation.

### 7.4.1 Prompting Baselines

**Zero-Shot (ZS)** The ZS setup used in our main experiment corresponds to the one from Section 7.3.1, in which the model is simply asked to produce a translation with no conditioning from examples or explanations towards the target translation style. We use this setting to establish a baseline style and translation quality performance for the models.

**Zero-Shot Explain (ZS-Exp)** Building upon the ZS setting, we experiment with a prompting strategy where LLMs are provided with detailed explanations of the most salient elements that characterize the desired translation style. We obtain such descriptions by prompting a capable proprietary model, GPT-4o (OpenAI, 2023), with 20 translations matching the desired style, asking it to synthesize a set of guidelines to match the examples. We evaluate two contrastive variants of this approach, providing GPT-4o with either MT examples ( $ZS-Exp_{HT}$ ) or alternative human translations ( $ZS-Exp_{PT}$ ) alongside examples matching the desired style, and asking to describe what characterizes

the latter compared to the former. To avoid data leakage, all generated explanations are manually reviewed to ensure they do not contain any verbatim content or direct excerpts from the input examples.<sup>5</sup> Tested models are then prompted with GPT-4o explanations in a ZS setting, to verify whether interpretable directives synthesized from a set of examples matching the desired behavior can produce reliable personalization results.

**Multi-Shot (MS)** Following Section 7.3.2’s findings, we adopt the same MS setup using 20 in-context translation examples matching the style of a target human translator ( $H_1$  or  $H_2$ ).

#### 7.4.2 Steering Baselines

We employ the Activation Addition (ActAdd) and Representation Fine-tuning (ReFT) methods introduced in Section 2.3.2 as baselines for comparing the effectiveness of our proposed method. For ActAdd, we employ the standard contrastive formulation by Rimsky et al. (2024) and Scalena et al. (2024) to extract two sets of style-relevant ( $\{z\}^+$ ) and default ( $\{z\}^-$ ) activations from a given model layer using 20 in-context examples demonstrating default behavior (MT) and the desired behavior ( $H_1$  or  $H_2$  translations), respectively. We then compute the average  $\Delta$  steering vector between the two sets of activations, scale it by a factor of  $\alpha = 2$ , which was found to be effective by previous research (Scalena et al., 2024), and apply it additively to the same model layer during inference. For ReFT, we apply learned interventions to the same personalization-relevant layers identified in Section 7.3.3 and limit confounding factors by tuning ReFT interventions with the set of 20 examples used for MS prompting.

#### 7.4.3 Contrastive SAE Steering

Given the SAE formulation we present in Section 2.3.2, our primary interest lies in the sparse latents  $h(z_l) \in \mathbb{R}^m$  learned by the SAE encoder, which were empirically found to capture monosemantic and interpretable properties of model inputs.

**Contrastive prompt setup** Given a set of paragraphs  $\mathcal{D}$  for a novel in the PAR3 dataset, each instance in it is a tuple:

$$\mathcal{D} = \{\langle s, H_1, H_2, \text{MT}_{\text{model}} \rangle\}$$

---

<sup>5</sup>Details on the prompt templates are in Section B.2.1.4.

with  $s$  being the non-English source sentence,  $H_1$  and  $H_2$  translations from two distinct human translators and  $MT_{model}$  the machine translation from the model under evaluation. Similar to previous methods, we employ a contrastive approach to extract SAE latents that are most active in the presence of the desired personalization style, while simultaneously controlling for more generic features that capture the generic properties of the task. We define two sets of contrastive prompts:

$$\mathcal{D}^+ = \{\langle s, e^+ \rangle\} \text{ and } \mathcal{D}^- = \{\langle s, e^- \rangle\}$$

capturing the personalized style of interest and baseline properties of the task, respectively. Similarly to the ZS-Exp setup from Section 7.4.1, we explore two  $\mathcal{D}^-$  configurations using either  $e^- = MT$  (SAE Cont.<sub>HT</sub>) or  $e^- = H_2$  (or  $H_1$ , if  $H_2$  is the personalization target; SAE Cont.<sub>PT</sub>) to assess the effect of baseline choice on steering effectiveness.

**Feature extraction** First, we gather activations  $z_l^+$  and  $z_l^-$  by prompting the model with inputs from the two contrastive sets  $\mathcal{D}^+$  and  $\mathcal{D}^-$ . Activations are extracted at the last prompt token position from its most informative layer, as identified in Section 7.3.3. Activations are then converted into sparse latent representations  $x^+ = h(z^+)$  and  $x^- = h(z^-)$ , with  $x^+, x^- \in \mathbb{R}^m$  by the SAE encoder. This procedure is repeated across 20 contrastive examples, resulting in two collections of SAE latent vectors for positive/negative examples:

$$\begin{aligned}\mathcal{X}^+ &= \{x_1^+, x_2^+, \dots, x_{20}^+\} \\ \mathcal{X}^- &= \{x_1^-, x_2^-, \dots, x_{20}^-\}\end{aligned}$$

**Relevance-based Feature Selection** To identify discriminative features for personalization in the large set of latents, we employ an information-theoretic approach adapted from Zhao et al. (2025). For each of the inputs, we identify the subset of size  $n < m$  that includes only the SAE active features, i.e., the latent dimensions for which the logit is greater than 0. We consider logit values in this subset as instances of a random variable  $X_i \in x$ , and calculate the mutual information  $I(X_i, Y)$  between each feature  $X_i$  and the target binary variable  $Y = \{+, -\}$  corresponding to the style of the provided examples (personalized or non-personalized). A higher  $I(X_i, Y)$  indicates that the  $i$ -th feature is more informative for discriminating between personalized and default inputs, and can hence be used for steering. A representative sample of 40 latents showing the highest mutual information scores for both personalized ( $\{X_i\}^+$ )

and non-personalized ( $\{X_i\}^-$ ) examples is selected using this procedure.<sup>6</sup> For every selected latent, we compute its expected logit when personalization is present or absent in provided examples, i.e.  $\mathbb{E}^+[X_i]$  and  $\mathbb{E}^-[X_i]$ .

**Inference-time intervention** Finally, activations are steered by setting selected latents to their expected value whenever their observed score is below (for the promoted personalized case) or above (for the demoted non-personalized case) the pre-computed average. Hence, in the SAE Cont.<sub>HT</sub> setting we enhance the features relevant to a target personalized style, e.g.  $\{X_i\}^{H_1}$  for  $H_1$ , and suppress the features  $\{X_i\}^{MT}$ , corresponding to the model’s default **MT**. In SAE Cont.<sub>PT</sub>, instead, we promote the same  $H_1$ -related latents while suppressing  $\{X_i\}^{H_2}$  to steer the model towards  $H_1$  personal style. Additionally, we modulate the magnitude of the resulting vector with an  $\alpha$  coefficient, which was found to play an essential role in steering effectiveness in previous research (Scalena et al., 2024; Ferrando et al., 2025). Algorithm 3 outlines the procedure for our proposed latent-based steering. It enhances features identified as relevant to personalization while simultaneously suppressing those negatively correlated with the task.

## 7.5 Experiments

### 7.5.1 Setup

**Model selection** We evaluate our methods on the same three models used for our preliminary evaluation of Section 7.3. Our selection is guided by the availability of open-source pre-trained SAEs, which can be computationally expensive to train otherwise. For Gemma models, we employ SAEs from the GemmaScope suite (Lieberum et al., 2024); for the Llama 3.1 model, we employ the SAE released by McGrath et al. (2024). GemmaScope SAEs are available for every model layer, enabling us to steer Gemma models on their most informative layers for the task, which we identified in Section 7.3.3. On the contrary, a single SAE for the 19th layer is available for Llama, hence limiting our evaluation of SAE steering and potentially producing sub-optimal steering results for that model.

**Metrics** We evaluate our approaches on a held-out test set sourced from the PAR3 dataset for personalization and output quality. For personalization, we use the classifiers described in Section 7.3.1. We define three submetrics employing the classifier

---

<sup>6</sup>By contrast, traditional SAE-based steering methods only employ features associated with the positive class (Chalnev et al., 2024; Arditì et al., 2024).

**Algorithm 3** Contrastive SAE Steering

---

**Require:** Input activation  $z$ , SAE model, target latents expected value  $\mathbb{E}^+[X_i]$ , contrast latents expected value  $\mathbb{E}^-[X_i]$ , steering coefficient  $\alpha$

**Ensure:** Steered activation  $z_{\text{new}}$

```

1: procedure CONTRASTIVESTEERING( $z$ , SAE,  $\mathbb{E}^+[X_i]$ ,  $\mathbb{E}^-[X_i]$ ,  $\alpha$ )
2:    $x = \text{SAE.encode}(z)$ 
3:    $m = \text{length}(x)$ 
4:   for  $i \leftarrow 1$  to  $m$  do
5:     if  $\mathbb{E}^+[X_i] > x[i]$  then
6:        $x[i] = \mathbb{E}^+[X_i]$ 
7:     end if
8:     if  $\mathbb{E}^-[X_i] < x[i]$  then
9:        $x[i] = \mathbb{E}^-[X_i]$ 
10:      end if
11:    end for
12:    $z_{\text{new}} = \alpha \cdot \text{SAE.decode}(x)$ 
13:   return  $z_{\text{new}}$ 
14: end procedure

```

---

probability distribution over the three classes ( $\text{MT}$ ,  $H_1$ ,  $H_2$ ) to better analyze different aspects of classifiers' predictions. First, we compute  $\mathbf{H}$  accuracy as the classifier's total probability assigned to human-like translations,  $p(H_1) + p(H_2)$ , thereby measuring the generic *human-like* style of the text. To measure personalization, we employ the personalization  $\mathbf{P}$ , corresponding only to the human translation currently selected as target ( $H_1$  or  $H_2$ ). Finally, the more stringent  $\mathbf{P}_{\text{flip}}$  metric measures the proportion of examples for which the applied conditioning procedure (either prompting or steering) causally influences the resulting classifier prediction, identifying examples for which the label flips from  $\text{MT}$  to the desired target.

To ensure that our interventions do not result in a degradation of overall translation quality, we also employ COMET<sup>7</sup> (Rei et al., 2020) using the personalized translation as reference.

**Quality-accuracy trade-off** We begin by verifying the optimal steering intensity  $\alpha$  for our SAE steering technique. We primarily focus on results from Gemma 2B, for

---

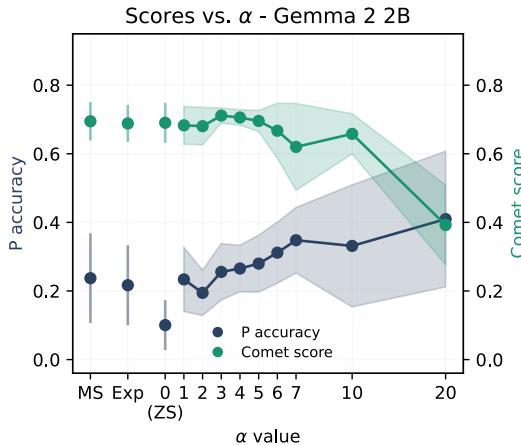
<sup>7</sup>Unbabel/wmt22-comet-da

Source (FR)	Cette somme vous sera restituée à votre sortie de prison, dit le juge. En attendant, vous êtes libres sous caution.		
H <sub>1</sub>	<i>This sum of money will be returned to you when you leave prison,' said the judge. In the meantime you are free on bail.</i>	-	-
ZS	<i>This amount will be returned to you upon your release from prison, the judge said. Meanwhile, you are free on bail.</i>	MT	0.79
SAE Cont. <sub>HT</sub>			
α = 5	<i>This sum will be repaid to you at your departure from prison, the judge declared. In the meantime, you are released under bond.</i>	MT	0.85
α = 50	<i>#echo all ought to goodness too was put moreover goodness behaving goodness goodness goodness goodness [...]</i>	H <sub>1</sub>	0.20
α = 150	<i>GURL callers goodness_Noneummy amidála Reportedaps Beaver-ishlor Decompiled_unset queer headquarters Headquarters tue ... [...]</i>	H <sub>1</sub>	0.23
Source (DE)	Nein – was Sie lieber haben, Herr Leutnant! Nur keine Zeremonien, es ist doch ganz einerlei.		
H <sub>2</sub>	<i>"No, no-whatever you would rather have, Lieutenant Hofmiller! Please don't stand on ceremony, it makes no difference to us."</i>	-	-
ZS	<i>No, whatever you prefer, Lieutenant! Just no ceremonies, it doesn't matter.</i>	MT	0.76
SAE Cont. <sub>HT</sub>			
α = 5	<i>No, anything at all you want, sir! Just don't make a fuss about it, it really doesn't matter.</i>	H <sub>2</sub>	0.79
α = 50	<i>"&gt;I Don't worry about that... I don't want a ceremony for this one. It's not important..."</i>	H <sub>2</sub>	0.46
α = 150	<i>IWhenInWhatItDonIf Sometimes AIs Celebrating ceremonies... Sosir please don't have parties ey' [...]</i>	H <sub>2</sub>	0.24

**Table 7.5 |** Examples from different languages being classified as Human when using extreme  $\alpha$  values.

which we ran a comprehensive sweep over all relevant hyperparameters.<sup>8</sup> Figure 7.3 illustrates the influence of  $\alpha$  on MT personalization accuracy and fluency averaged across all translators for all tested languages. For values of  $\alpha \leq 3$ , performance remains close to that of the MS baseline, indicating that the contrastive method is effectively isolating latents associated with human-like style. As  $\alpha$  increases, performance generally exceeds the MS approach, achieving greater control and flexibility in guiding the model's output with minimal impact on translation quality. However, for  $\alpha \geq 10$ , we observe a notable degradation in COMET, suggesting an important drop in translation fluency. Table 7.5

<sup>8</sup>Larger models were evaluated using a subset of the best-performing configurations. Details in Section B.2.2.



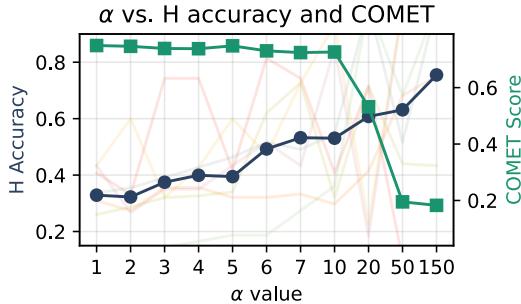
**Figure 7.3 |** Personalization  $\mathbf{P}$  and COMET across various steering intensity  $\alpha$  for SAE Cont.<sub>HT</sub> on Gemma 2 2B. The performance of zero-shot multi-shot (MS), zero-shot explain (Exp) and zero-shot ( $ZS = 0$ ) baselines is also reported.

shows some examples of models generating output aligned with the Human translator according to the classifier, but with a low COMET score corresponding to an almost unreadable output due to extreme  $\alpha$  values. False positive classifications in such settings suggest that steering methods and classifiers are aligned with potentially spurious stylistic features, which are not necessarily indicative of high-quality translations. We leave the investigation of these spurious features to future work, focusing here on the trade-off between personalization and translation quality.

Following Ferrando et al. (2025), which also employ SAEs for steering, we experiment with very high alpha values (up to 150), finding the classifier’s  $\mathbf{H}$  accuracy approaching 100% for some languages. While this indicates that the contrastive steering is aggressively optimizing toward classifier preferences (Figure 7.4), the consequent drop in COMET scores reveals a steep decline in translation quality, often resulting in incoherent or nonsensical generations from a human perspective. Ultimately, we identify  $\alpha = 5$  as an appropriate steering intensity to balance personalization and fluency, and employ it for our main evaluation.

### 7.5.2 Results and Discussion

Table 7.6 presents performances of tested models across prompting and steering setups, averaged across all languages and personalization targets ( $\mathbf{H}_1$  and  $\mathbf{H}_2$  for each language).



**Figure 7.4 |** COMET and H accuracy across  $\alpha$  steering intensity values for Gemma 2 2B, showing a major drop in translation quality for very high intensities ( $\alpha \geq 50$ ).

We find that our SAE  $\text{Cont}_{\text{HT}}$  and SAE  $\text{Cont}_{\text{PT}}$  methods generally achieve the best trade-off between personalization accuracy and translation quality, especially for the smaller Gemma 2 2B model. This could be due to the larger models’ superior ability to incorporate in-context information naturally, reducing the relative benefit of explicit steering. Comparing the two contrastive setups (HT and PT) for the ZS-Exp and SAE Cont. methods, we find that using different human demonstrations as a contrastive baseline in PT generally produces better results for larger models. As for general performance, we conjecture this could be due to the larger models’ improved ability to disentangle personalization-critical factors without explicit guidance. For the smaller Gemma 2 2B, the difference between the two approaches is minimal, suggesting that the model cannot fully exploit the differences between the examples.

### Do SAE Steering and MS Prompting Impact Activations in a Similar Way?

Since SAE-based approaches perform on par or better than MS, we set out to investigate whether the two methods result in a similar impact on model representations. We collect the modified activations  $z_{\text{steer}}$  obtained from the SAE  $\text{Cont}_{\text{HT}}$  steering setting and evaluate them using the probing classifier trained on MS-conditioned activations, as introduced in Section 7.3.3, for detecting personalization information. Table 7.7 shows probe accuracy in detecting the positive impact of SAE steering across the three possible outcomes of the steering procedure. We find that the probe trained on the SAE layer effectively distinguishes between activations corresponding to successful and unsuccessful SAE steering, despite having been exposed only to MS conditioning during training. This includes both instances where the classifier prediction is flipped after steering ( $\text{MT} \rightarrow \text{H}$ ) and settings where the conditioning fails ( $\text{MT} \rightarrow \text{MT}$ ). In settings where the original output already matches human style ( $\text{H} \rightarrow \text{H}'$ ), the probe

	Gemma 2 2B				Gemma 2 9B				LLaMA 3.1 8B			
	H	P	$P_{\text{flip}}$	$\textcolor{teal}{f}$	H	P	$P_{\text{flip}}$	$\textcolor{teal}{f}$	H	P	$P_{\text{flip}}$	$\textcolor{teal}{f}$
ZS	0.21	0.10	0.05	0.69	0.15	0.08	0.04	0.71	0.24	0.08	0.05	0.70
ZS-Exp. <sub>HT</sub>	0.30	0.22	0.16	0.68	0.41	0.22	0.18	0.72	0.56	0.23	0.21	0.69
ZS-Exp. <sub>PT</sub>	–	0.20	0.14	0.69	–	0.23	0.19	<b>0.73</b>	–	0.30	0.26	0.70
MS	0.37	0.24	0.16	0.69	<b>0.48</b>	0.31	0.27	<b>0.73</b>	0.58	0.32	<b>0.28</b>	<b>0.73</b>
ActAdd	0.27	0.22	0.12	0.67	0.32	0.24	0.20	0.70	0.55	0.36	<b>0.28</b>	0.70
ReFT	0.31	0.22	0.18	<b>0.70</b>	0.46	0.34	0.27	0.67	0.53	<b>0.38</b>	0.26	0.70
SAE Cont. <sub>HT</sub>	<b>0.39</b>	<b>0.27</b>	<b>0.19</b>	<b>0.70</b>	0.46	0.33	<b>0.29</b>	0.72	<b>0.59</b>	0.31	0.27	0.72
SAE Cont. <sub>PT</sub>	–	<b>0.27</b>	0.18	0.69	–	<b>0.35</b>	<b>0.29</b>	<b>0.73</b>	–	0.33	<b>0.28</b>	0.72

**Table 7.6 |** Averaged metric scores across all tested languages (per-language breakdown in appendix). **H:** human style accuracy, i.e.  $p(\mathbf{H}_1) + p(\mathbf{H}_2)$ . **P:** personalization accuracy  $p(\mathbf{H}_x)$  for the target style.  **$P_{\text{flip}}$ :** Proportion of segments for which steering has a causal impact on personalization.  $\alpha = 5$  is used for SAE Cont. results.

	MT → H*	MT → MT	H* → H*
Gemma 2 2B	$0.94 \pm 0.01$	$0.07 \pm 0.02$	$0.72 \pm 0.15$
Gemma 2 9B	$0.93 \pm 0.02$	$0.12 \pm 0.10$	$0.68 \pm 0.19$

**Table 7.7 |** Probing accuracy on ContHT-steered activations averaged across languages and  $\mathbf{H}_1/\mathbf{H}_2$  translators. Probes trained on MS activations reliably detect the impact of SAE Cont. steering over model generations, suggesting similar mechanisms.

obtains lower accuracy with broader confidence intervals, denoting higher uncertainty. These findings suggest that the **SAE’s latents we extract through our contrastive method are meaningfully connected to the stylistic patterns embedded in the multi-shot examples**, providing evidence that our intervention influences the internal representations of the model, aligning them to the natural effect of the MS approach.

## 7.6 Limitations

While we demonstrate the potential of steering LLMs for MT personalization using sparse autoencoders, we acknowledge several limitations.

Firstly, the generalizability of our findings is constrained by the scope of our experiments. We focused on literary translation from seven specific source languages into English and evaluated three LLMs of relatively small size. Consequently, the observed effectiveness of SAE-based steering and the identified optimal layers for intervention may not directly transfer to other language pairs, significantly different model archi-

lectures or sizes, or distinct domains beyond literary texts. Further research is needed to assess the robustness of our approach across a broader range of linguistic and modeling contexts.

Secondly, the computational overhead associated with sparse autoencoders presents a practical challenge. Although we utilized pre-trained SAEs in our study, the initial training of these components is resource-intensive. This could limit the accessibility and scalability of our proposed method, particularly for researchers or practitioners with limited computational resources or when frequent retraining for new models or tasks is required. The current availability of pre-trained SAEs also restricts model choice, as seen with the Llama 3.1 8B model, where an SAE was only available for a potentially sub-optimal layer.

Finally, our investigation primarily focused on downstream performance and the impact of various personalization strategies on model representations. However, we did not pursue a mechanistic understanding of the “personalization circuits” within the LLMs. Future work could adopt a more fine-grained mechanistic interpretability approach to study how specific SAE latents or combinations thereof encode and manipulate nuanced stylistic features, thereby providing deeper insights into the underlying processes of LLM personalization.

## 7.7 Conclusion

We conducted a broad evaluation of various prompting and steering approaches for personalizing LLM-generated translations. Our evaluation targets a practical, real-world application of literary translation and addresses the underexplored challenge of steering LLM generations in a linguistically rich and stylistically sensitive domain. Through comprehensive evaluation across multiple languages, novels, and models, we demonstrate that our proposed SAE-based approach outperforms prompting and alternative steering techniques.

Although faithfully replicating individual human translation styles remains a highly challenging task, our approach achieves strong alignment with human translation quality, as reflected in both general human-likeness and translator-specific personalization metrics. These results highlight the method’s robustness and its potential to support high-fidelity translation workflows in real-world settings. Concretely, these results have important implications in the development of personalized MT systems based on LLMs. In particular, the notable effectiveness of our proposed approach on smaller models might enable MT customization when few examples are available, facilitating further

research on how personalization information is encoded and produced by language models. Despite their effectiveness, the interpretability of the learned SAE latents and their potential use with larger LLMs—where increased capacity may further enhance the precision and fluency of personalized translations—remain open questions for future investigation.

The success of SAE-based steering for personalized translation highlights the effectiveness of internal-based interventions for controlling model generation. However, the practical deployment of such approaches requires a careful evaluation of their impact on users' trust and behaviors. The first two parts of this thesis focused on developing methods for analyzing and steering model generation. The third and final part, beginning with the next chapter, focuses instead on the *users* of machine translation systems, specifically professional post-editors and translators, to explore how their interactions with machine-translated content are shaped by factors such as language similarity and translation quality. Finally, we investigate whether these interactions can be improved using the trove of information available from the inner workings of MT models.



## PART III

# Interpretability in Human Translation Workflows



# CHAPTER 8

---

## Machine Translation Post-editing for Typologically Diverse Languages

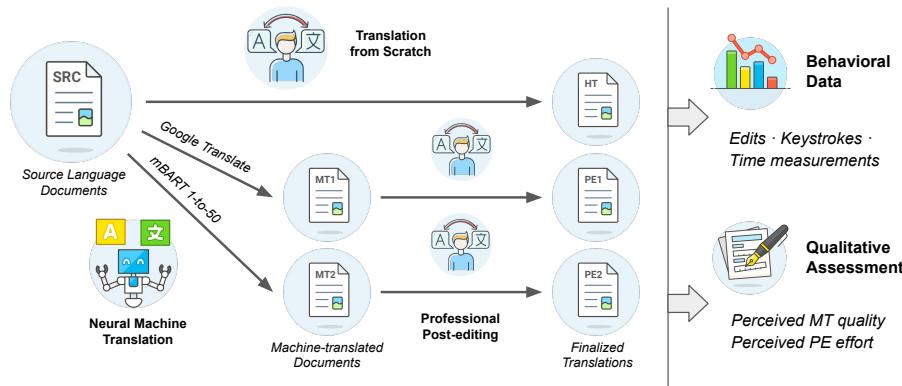
### Chapter Summary

This chapter begins our investigation into the application of interpretability methods in user-facing translation settings. As an initial step in this direction, we introduce DivEMT, the first publicly available post-editing dataset spanning six typologically diverse target languages. We evaluate the impact of MT quality and translation directions on post-editing effectiveness in a controlled setup involving 18 professional editors through comprehensive behavioral logging of edits, keystrokes, timing, and pauses. While we find that post-editing machine translation is consistently faster than translation from scratch, our results show significant disparities across languages with different typological relationships to English, even when controlling for system architecture and data size, highlighting the need for tailored approaches in MT for diverse languages.

This chapter is adapted from the paper *DivEMT: Neural Machine Translation Post-Editing Effort Across Typologically Diverse Languages* (Sarti et al., 2022).

*Language was just difference. A thousand different ways of seeing, of moving through the world. No, a thousand worlds within one. And translation, a necessary endeavor however futile, to move between them.*

– Rebecca F. Kuang, *Babel* (2022)



**Figure 8.1 |** The DivEMT data collection process. For every English source document, 18 professional translators are tasked with translating it from scratch (HT) or post-editing NMT systems’ outputs (PE<sub>1</sub>/PE<sub>2</sub>) into six typologically diverse target languages. Behavioral data and qualitative assessments are collected during and after the process, respectively.

### 8.1 Introduction

Recent advances in neural language modeling and multilingual training have led to the widespread adoption of machine translation (MT) technologies across an unprecedented range of languages worldwide. While the benefits of state-of-the-art MT for cross-lingual information access are undisputed (Gene, 2021), its usefulness as an aid to professional translators varies considerably across domains, subjects and language combinations (Zouhar et al., 2021b). In the last decade, the MT community has been including an increasing number of languages in its automatic and human evaluation efforts (Bojar et al., 2013; Barrault et al., 2021). However, the results of these evaluations are typically not directly comparable across different language pairs for several reasons. First, reference-based automatic quality metrics are hardly comparable across different target languages (Bugliarello et al., 2020). Second, human judgments are collected independently for different language pairs, making their cross-lingual comparison vulnerable to confounding factors such as tested domains and training data sizes. Similarly, recent work on NMT post-editing efficiency has focused on specific language pairs such as English-Czech (Zouhar et al., 2021b), German-Italian, German-French (Läubli et al., 2019) and English-Hindi (Ahsan et al., 2021). However, a controlled comparison across a set of typologically diverse languages is still lacking.

In this chapter, we conduct an initial assessment of the usefulness of state-of-the-art

NMT in professional translation with a strictly controlled cross-language setup (Figure 8.1). Specifically, professionals were asked to translate the same English documents into six typologically distinct languages—Arabic, Dutch, Italian, Turkish, Ukrainian, and Vietnamese—using the same platform and guidelines. Three **translation modalities** were adopted: human translation from scratch (HT), post-editing of Google Translate’s translation (PE<sub>1</sub>), and post-editing of mBART-50’s translation (PE<sub>2</sub>), the latter being a state-of-the-art open-source, multilingual NMT system. In addition to post-editing results, subjects’ fine-grained editing behavior, including keystrokes and time information, was logged to measure productivity and effort across languages, systems and translation modalities. Finally, translators were asked to complete a qualitative assessment regarding their perceptions of MT quality and post-editing effort. The resulting DivEMT dataset, to our best knowledge, is the first public resource that allows a direct comparison of professional translators’ productivity and fine-grained editing information across a set of typologically diverse languages. All collected data are publicly released<sup>1</sup> alongside this paper to foster further research in the language- and system-dependent nature of NMT advances in real-world translation scenarios.

## 8.2 Related Work

**Cross-lingual MT Evaluation** Before the advent of NMT, Birch et al. (2008) studied how various language properties affected the quality of Statistical MT (SMT) across a sizeable sample of European language pairs. The comparison, however, was solely based on BLEU, which is not directly comparable across different target languages (Bugliarello et al., 2020). Recent work on neural models introduced more principled ways to measure the intrinsic difficulty of language-modeling (Gerz et al., 2018; Cotterell et al., 2018; Mielke et al., 2019) and machine-translating (Bugliarello et al., 2020; Bisazza et al., 2021) different languages. However, reliably achieving this without human evaluation remains an open research question. Concurrently to our research, Licht et al. (2022) proposed a new human evaluation protocol to improve consistency in cross-lingual MT quality assessment.

**Post-editing NMT** Recent work highlighted the productivity gains driven by NMT post-editing on a broader array of languages that were previously challenging for MT, such as English-Dutch (Daems et al., 2017a), English-Hindi (Ahsan et al., 2021), English-Greek (Stasimioti and Sosoni, 2020), English-Finnish and English-Swedish (Koponen et al., 2020), all showing a considerable variance among language

---

<sup>1</sup> [GroNLP/divemt](#)

pairs and subjects. Interestingly, Zouhar et al. (2021b) found that NMT post-editing speed was comparable to translation from scratch in English–Czech, and highlighted a disconnect between moderate increases in automatic MT quality metrics and improved post-editing productivity. In summary, research on post-editing NMT generally reports increased fluency and output quality; however, productivity gains are hardly generalizable across language pairs and domains. Importantly, to our knowledge, no previous work has studied NMT post-editing over a set of typologically different languages while controlling for the effects of content types and domains, NMT engines, and translation interfaces.

### 8.3 The DivEMT Dataset

DivEMT’s primary purpose is to assess the usefulness of state-of-the-art NMT for professional translators and to study how this usefulness varies across target languages with different typological properties. We present below our data collection setup, which strikes a balance between simulating a realistic professional translation workflow and maximizing the comparability of results across languages.

#### 8.3.1 Subjects and Task Scheduling

To control for the effect of individual translators’ preferences and styles, we involve a total of 18 subjects (three per target language). During the experiment, each subject receives a series of short *documents* (3 to 5 sentences each) where the source text is presented in isolation (HT) or alongside a translation proposal produced by one of the NMT systems (PE<sub>1</sub>, PE<sub>2</sub>). The experiment comprises two phases: during the **warm-up phase** a set of 5 documents is translated by all subjects following the same, randomly sampled sequence of modalities (HT, PE<sub>1</sub> or PE<sub>2</sub>). This phase allows the subjects to become accustomed to the setup and enables us to identify potential issues in the logged behavioral data before proceeding.<sup>2</sup> In the **main collection phase**, each subject is asked to translate documents in a pseudo-random sequence of modalities. This time, however, the sequence is different for each translator and chosen so that each document gets translated in all three modalities. This allows us to measure translation productivity independently from the subject’s productivity and document-specific difficulties.

Table 8.1 shows an example of the adopted modality scheduling. The modality of document docM<sub>i</sub> for translator T<sub>j</sub> in the main task is picked randomly among the two modalities that were not seen by the same translator for docM<sub>i-1</sub>, enforcing consecutive

---

<sup>2</sup>Warm-up data are excluded from the analysis of Section 8.4.

		T <sub>1</sub>	T <sub>2</sub>	T <sub>3</sub>
warm-up	docW <sub>1</sub>	HT	HT	HT
	docW <sub>2</sub>	PE <sub>1</sub>	PE <sub>1</sub>	PE <sub>1</sub>
	...			
	docW <sub>N</sub>	PE <sub>2</sub>	PE <sub>2</sub>	PE <sub>2</sub>
main	docM <sub>1</sub>	HT	PE <sub>1</sub>	PE <sub>2</sub>
	docM <sub>2</sub>	PE <sub>2</sub>	HT	PE <sub>1</sub>
	docM <sub>3</sub>	HT	PE <sub>2</sub>	PE <sub>1</sub>
	...			
	docM <sub>N</sub>	PE <sub>2</sub>	PE <sub>1</sub>	HT

**Table 8.1 |** Modality scheduling overview. For each language, each subject ( $T_i$ ) works with a pseudo-random sequence of modalities (HT, PE<sub>1</sub>, PE<sub>2</sub>). For the warm-up task ( $N=5$ ), all translators are provided with the same documents in the same modalities. For the main task ( $N=107$ ), each translator is assigned a modality at random. Each document is translated once for every modality. The same procedure is repeated independently for all the languages.

documents given to the same translator to be assigned different modalities to avoid periodicity in repetition and enable same-language comparisons. Importantly, although all three modes were collected for every document, we did not enforce mode consistency across the same translator identifier across languages (i.e. T<sub>1</sub> for Italian does not have the same sequence of modalities of translator T<sub>1</sub> in Arabic, for example). For this reason, individual subjects are not directly comparable across languages. This is relevant since comparable editing behavior should be attributed to similar personal preferences rather than an identical modality assignment of the same sentences. Despite modality scheduling, we have no guarantees that translators consistently follow the order of documents presented in PET, and thus possibly operate on documents assigned to the same modality consecutively. However, this possibility reduces to random guessing due to a lack of any identifying information related to the modality until the document is entered for editing. The sequence of modalities for the warmup task is fixed and is: HT, PE<sub>2</sub>, PE<sub>1</sub>, HT, PE<sub>2</sub>.

As productivity and other behavioral metrics can only be estimated with a sizable sample, we prioritize the number of documents over the number of subjects per language during budget allocation. In future analyses, a larger set of post-edited documents would also provide more insight into the error type distribution of NMT systems across different language pairs.

All subjects are professional translators with at least 3 years of professional experience,

including at least 1 year of post-editing experience, and strong proficiency in CAT tools.<sup>3</sup> Translators were provided with links to the source articles to facilitate contextualization, were asked to produce translations of publishable quality and were instructed not to use any external MT engine to produce their translations. Assessing the final quality of the post-edited material is out of the scope of the current study, although we realize that this is an important consideration to assess usability in a professional context.<sup>4</sup>

### 8.3.2 Choice of Source Texts

The selected documents represent a subset of the FLORES-101 benchmark (Goyal et al., 2022) consisting of sentences taken from English Wikipedia, and covering a mix of topics and domains.<sup>5</sup> While professional translators generally specialize in one or a few domains, we opt for a mixed-domain dataset to minimize domain adaptation efforts by the subjects and maximize the generalizability of our results. Importantly, FLORES-101 includes high-quality human translations into 101 languages, which enables the automatic estimation of NMT quality and the discarding of excessively low-scoring models or language pairs before our experiment. FLORES-101 also provides valuable metadata, e.g. source URL, which allows us to ensure the absence of public translations of the selected contents, which could be leveraged by translators and compromise the validity of our setup. The documents used for our study are fragments of contiguous sentences extracted from Wikipedia articles that compose the original FLORES-101 corpus. Even if small, the context provided by document structure allows us to simulate a more realistic translation workflow if compared to out-of-context sentences.

Based on our available budget, we selected 112 English documents from the *devtest* portion of FLORES-101, corresponding to 450 sentences and 9,626 words. More details on the data selection process are provided in Section C.1.3.

### 8.3.3 Choice of Languages

Training data is one of the most important factors in determining the quality of an NMT system. Unfortunately, using strictly comparable or multi-parallel datasets, such as Europarl (Koehn, 2005) or the Bible corpus (Mayer and Cysouw, 2014), would dramatically restrict the diversity of languages available to our study or imply prohibitively low translation quality on general-domain text. In order to minimize the effect of

---

<sup>3</sup>Additional subjects' details are available in Section C.1.1.

<sup>4</sup>A summary of our translation guidelines is provided in Section C.1.2.

<sup>5</sup>We use a balanced sample of articles sourced from WikiNews, WikiVoyage and WikiBooks.

Genus:Family		$d_{syn}$	Morphology	MSP	TTR	Script
ENG	Indo-European:Germanic	–	Fusional	1.17	0.28	latin
ARA	Afro-Asiatic:Semitic	0.57	Introflexive	1.67	0.46	arabic
NLD	Indo-European:Germanic	0.49	Fusional	1.16	0.28	latin
ITA	Indo-European:Romance	0.51	Fusional	1.30	0.30	latin
TUR	Altaic:Turkic	0.70	Agglutinative	2.28	0.50	latin
UKR	Indo-European:Slavic	0.51	Fusional	1.42	0.47	cyrillic
VIE	Austro-Asiatic:VietMuong	0.57	Isolating	1.00	0.12	latin

**Table 8.2 |** Typological diversity of our language sample.  $d_{syn}$ : Syntactic distance w.r.t. English Lin et al. (2019). MSP: Mean size of paradigm, from Çöltekin and Rama (2023). TTR: Type-token ratio measured on FLORES-101. Shading indicates genetic/syntactic relatedness to English and morphological complexity/lexical richness.

training data disparity while maximizing language diversity, we choose representatives of six different language families for which comparable amounts of training data are available in our open-source model, namely **Arabic**, **Dutch**, **Italian**, **Turkish**, **Ukrainian**, and **Vietnamese**. As shown in Table 8.2, our language sample exhibits a good diversity in terms of language family, relatedness to English, type of morphological system, morphological complexity, measured by the mean size of paradigm (MSP, Xanthos et al., 2011), and script. We also report the type-token ratio (TTR), the only language property found to correlate significantly with translation difficulty in a sample of European languages (Bugliarello et al., 2020). While the amount of language-specific parallel sentence pairs used for the multilingual fine-tuning of mBART-50 varies widely ( $4K < N < 45M$ ), all our selected language pairs fall within the  $100K\text{--}250K$  range (mid-resourced, see Table 8.3), enabling a fair cross-lingual performance comparison.

#### 8.3.4 Choice of MT Systems

While most of the best-performing general-domain NMT systems are commercial, experiments based on such systems are not replicable, as their backends are silently updated over time. Moreover, without knowing the exact training specifics, we cannot attribute differences in the cross-lingual results to intrinsic language properties. We balance these observations by including two NMT systems in our study: **Google Translate** (GTrans)<sup>6</sup> as a representative of commercial quality, and **mBART-50 one-to-Many**<sup>7</sup>

<sup>6</sup>Evaluation performed in October 2021.

<sup>7</sup>`mbart-large-50-one-to-many`

	Google Translate (PE <sub>1</sub> )	mBART-50 (PE <sub>2</sub> )	# Pairs
ARA	<b>34.1 / 65.6 / .737</b>	17.0 / 48.5 / .452	226K
NLD	<b>29.1 / 60.0 / .667</b>	22.6 / 53.9 / .532	226K
ITA	<b>32.8 / 61.4 / .781</b>	24.4 / 54.7 / .648	233K
TUR	<b>35.0 / 65.5 / 1.00</b>	18.8 / 52.7 / .755	204K
UKR	<b>31.1 / 59.8 / .758</b>	21.9 / 50.7 / .587	104K
VIE	<b>45.1 / 61.9 / .724</b>	34.7 / 54.0 / .608	127K

**Table 8.3 |** MT quality of the selected NMT systems for English-to-Target translation on the full FLORES-101 devtest split, in BLEU / ChrF / COMET format. Best scores are highlighted in **bold**. We report the number of sentence pairs used for mBART-50 multilingual fine-tuning by Tang et al. (2021).

(Tang et al., 2021) as a representative of state-of-the-art open-source multilingual NMT technology. The original multilingual BART model (Liu et al., 2020) is an encoder-decoder transformer model pre-trained on monolingual documents in 25 languages. Tang et al. (2021) extend mBART by further pre-training on 25 new languages and performing *multilingual translation fine-tuning* for the full set of 50 languages, producing three configurations of multilingual NMT models: many-to-one, one-to-many, and many-to-many. Our choice of mBART-50 is primarily motivated by its manageable size, good performance across the set of evaluated languages (see Table 8.3), and its adoption for other NMT studies (Liu et al., 2021) and post-editing evaluations (Fomicheva et al., 2022b). Although mBART-50 performances are usually comparable or slightly worse than those of tested bilingual NMT models,<sup>8</sup> using a multilingual model allows us to evaluate the downstream effectiveness of a single, unified system trained on pairs evenly distributed across tested languages. Finally, adopting two systems with marked differences in automatic evaluation scores allows us to estimate how a significant increase in metrics such as BLEU, ChrF and comet (Papineni et al., 2002; Popović, 2015; Rei et al., 2020) impacts downstream productivity across languages in a realistic post-editing scenario.

### 8.3.5 Translation Platform and Collected Data

Translators were asked to use PET (Aziz et al., 2012), a computer-assisted translation tool that supports both translating from scratch and post-editing. This tool was chosen because (i) it logs information about the post-editing process, which we use to assess

<sup>8</sup>See Section C.1.4 for automatic MT quality results by five different models over a larger set of 10 target languages.

effort (see Section 8.4); and (ii) it is a mature research-oriented tool that has been successfully used in several previous studies (Koponen et al., 2012; Toral et al., 2018b), and we modify it slightly to support right-to-left languages like Arabic. Using PET, we collect three types of data:

- **Resulting translations** produced by translators in either HT or PE modes, constituting a multilingual corpus with one source text and 18 translations (one per language-modality combination) exemplified in Table 8.4.
- **Behavioral data** for translated sentences, including editing time, amount and type of keystrokes (content, navigation, erase, etc.), and number and duration of pauses above 300/1000 milliseconds (Lacruz et al., 2014).
- **Pre- and post-task questionnaire.** The former focuses on demographics, education, and work experience with translation and post-editing. The latter elicits subjective assessments of post-editing quality, effort and enjoyability compared to translating from scratch.

## 8.4 Post-Editing Effort Across Languages

In this section, we use the DivEMT dataset to quantify the post-editing effort of professional translators across our diverse set of target languages. We consider two main objective indicators of editing effort: *temporal measurements* (and related productivity gains) and *post-editing rates*, measured by the Human-targeted Translation Edit Rate (HTER, Snover et al. (2006)). Finally, we assess the subjective perception of PE gains by examining the post-task questionnaires. We reiterate that all scores in this section are computed on the same set of source sentences for all languages, resulting in a faithful cross-lingual comparison of post-editing effort thanks to DivEMT’s controlled setup.

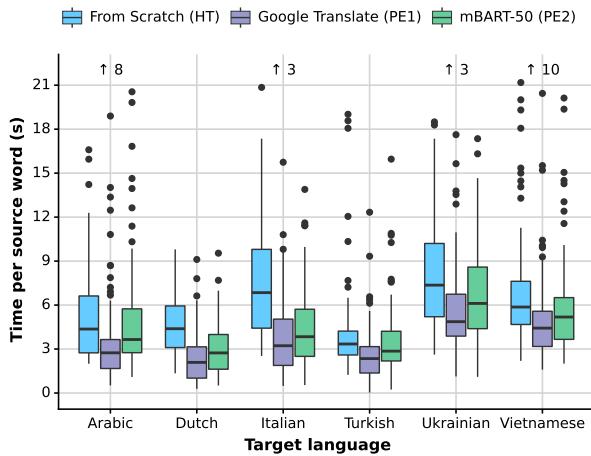
### 8.4.1 Temporal Effort and Productivity Gains

We begin by comparing *task time* (seconds per processed source word) across languages and modalities. For this purpose, edit times are computed for every document in every language without considering the presence of multiple translators for every language. As shown in Figure 8.2, translation time varies considerably across languages even when no MT system is involved (HT), suggesting an intrinsic variability in translation complexity for different subjects and language pairs. Indeed, for the HT modality, the time required for the “slowest” target languages (Italian, Ukrainian) is roughly twice that of the “fastest” one (Turkish). This pattern cannot be easily explained and contrasts

ENG	SRC	Inland waterways can be a good theme to base a holiday around.
ARA	HT	عَكْنُ أَنْ تَكُونُ الْمَرَاتِ الْمَائِيَّةِ الدَّاخِلِيَّةِ خَيَارًا جَيِّدًا لِتَخْطِيطِ عَطْلَةِ حَوْلَهَا.
	MT	يمكن أن تكون السكك الحديدية الداخلية موضوعاً جيداً لإقامة عطلة حول.
	PE	قد تكون المرات المائية الداخلية مكاناً جيداً لقضاء عطلة حولها.
NLD	HT	Binnenlandse waterwegen kunnen een goed thema zijn voor een vakantie.
	MT	Binnenwaterwegen kunnen een goed thema zijn om een vakantie rond te zetten.
	PE	Binnenwaterwegen kunnen een goed thema zijn om een vakantie rond te organiseren.
ITA	HT	I corsi d'acqua dell'entroterra possono essere un ottimo punto di partenza da cui organizzare una vacanza.
	MT	I corsi d'acqua interni possono essere un buon tema per fondare una vacanza.
	PE	I corsi d'acqua interni possono essere un buon tema su cui basare una vacanza.
TUR	HT	İç bölgelerdeki su yolları, tatil planı için iyi bir tema olabilir.
	MT	İç suyolları, tatil için uygun bir tema olabilir.
	PE	İç sular tatil için uygun bir tema olabilir.
UKR	HT	Можна спланувати вихідні, взявшись за основу подорож внутрішніми водними шляхами.
	MT	Водні шляхи можуть бути хорошим об'єктом для базування відпочинку навколо.
	PE	Місцевість навколо внутрішніх водних шляхів може бути гарним вибором для організації відпочинку.
VIE	HT	Du lịch trên sông có thể là một lựa chọn phù hợp cho kỳ nghỉ.
	MT	Các tuyến nước nội địa có thể là một chủ đề tốt để xây dựng một kì nghỉ.
	PE	Du lịch bằng đường thủy nội địa là một ý tưởng nghỉ dưỡng không tồi.

**Table 8.4 |** A DivEMT corpus entry, including the English source (SRC), its translation from scratch (HT), the MT output of mBART-50 (MT) and its post-edited version (PE) for all languages. We highlight insertions, deletions, substitutions and shifts computed with Tercom Snover et al. (2006). Full examples available in Table C.5.

with factors commonly tied to MT complexity, such as source-target morphological richness and language relatedness (Birch et al., 2008; Belinkov et al., 2017). On the other hand, we find that the relation  $PE_1 < PE_2 < HT$  (where  $PE_1$  is the fastest,  $PE_2$  has a medium speed, and  $HT$  is the slowest) holds for all the evaluated languages.



**Figure 8.2 |** Temporal effort across languages and translation modalities, measured in seconds per processed source word. Each point represents a document, with higher scores denoting slower editing. ↑: amount of data points per language not shown in the plot.

For a measure of productivity gains that is easier to interpret and more in line with translation industry practices, we turn to *productivity* expressed in source words processed per minute and compute the *speed-up* induced by the two post-editing modalities over translating from scratch ( $\Delta HT$ ). Table 8.5 presents our results. **Across systems**, we find that *large* differences among automatic MT quality metrics indeed reflect post-editing effort, suggesting a nuanced picture that complements the findings of Zouhar et al. (2021b). While post-editing time gains were observed to quickly saturate for slight changes in high-quality MT, we find that moving from medium-quality to high-quality MT yields meaningful productivity improvements across most evaluated languages. **Across languages**, too, the magnitude of productivity gains ranges widely, from doubling in some languages (Dutch PE<sub>1</sub>, Italian PE<sub>1</sub> and PE<sub>2</sub>) to only about 10% (Arabic, Turkish and Ukrainian PE<sub>2</sub>). When only considering the better-performing system (PE<sub>1</sub>), post-editing remains clearly beneficial in all languages despite the high variability in  $\Delta HT$  scores. Results are more nuanced for the open-source system (PE<sub>2</sub>), with three out of six languages displaying only marginal gains (<15% in Arabic, Turkish

	PROD ↑			ΔHT ↑	
	HT	PE <sub>1</sub>	PE <sub>2</sub>	PE <sub>1</sub>	PE <sub>2</sub>
ARA	13.1	21.7	16.3	+84%	+10%
NLD	13.6	28.7	21.7	+119%	+61%
ITA	8.8	18.6	15.6	+96%	+95%
TUR	17.9	25.5	21.0	+34%	+12%
UKR	8.0	12.3	9.8	+71%	+14%
VIE	10.2	13.0	11.1	+32%	+23%

**Table 8.5 |** Median productivity (PROD, # processed source words per minute) and median % post-editing speedup (ΔHT) for all analyzed languages and modalities. Arrows denote the direction of improvement.

and Ukrainian). Despite its overall inferior performance, mBART-50 (PE<sub>2</sub>) is the only system that enables a fair comparison across languages (in terms of training data size and architecture, see Section 8.3.4). Interestingly, when focusing on the productivity gains achieved by this system, factors such as language relatedness and morphological complexity become relevant. Specifically, Italian (+95%), Dutch (+61%) and Ukrainian (+14%) are genetically and syntactically related to English, but Ukrainian has a richer morphology (see Table 8.2). On the other hand, Vietnamese (+23%), Turkish (+12%) and Arabic (+10%) all belong to different families. However, Vietnamese is isolating (little to no morphology), while Turkish and Arabic have rich morphological systems (respectively agglutinative and introflexive, the latter of which is especially problematic for subword segmentation, Amrhein and Sennrich (2021)). Other differences, however, are more difficult to explain. For instance, Dutch is closely related to English and has a simpler morphology than Italian, but its productivity gain with mBART-50 is lower (61% vs 95%). This finding is accompanied by an important gap in BLEU and COMET scores achieved by mBART-50 on the two languages (22.6 vs 24.4 BLEU and 0.532 vs 0.648 COMET for Dutch vs Italian, resp.), which cannot be explained by training data size.

In summary, our findings confirm the overall positive impact of NMT post-editing on translation productivity observed in previous PE studies. However, we note that *the magnitude of this impact is highly variable across systems and languages*, with inter-subject variability also playing an important role, in line with previous studies (Koponen et al., 2020) (see Section 8.5 for more details). The small size of our language sample does not allow us to draw direct causal links between specific typological properties and post-editing efficiency. That said, we believe these results have important implications for the claimed ‘universality’ of current state-of-the-art MT and NLP systems, primarily

based on the transformer architecture (Vaswani et al., 2017) and BPE-style subword segmentation techniques (Sennrich et al., 2016b).

#### 8.4.1.1 Modeling Temporal Effort

Given the high variability among translators, segments and translation modalities, we assess the validity of our observations via statistical analysis of temporal effort using a linear mixed-effects regression model (LMER, Lindstrom and Bates (1988)), following Green et al. (2013) and Toral et al. (2018b). Linear Mixed Effects models (LMER) are used for regression analyses involving dependent data, such as longitudinal studies with multiple observations per subject. We fit our model on  $n = 7434$  instances, corresponding to 413 sentences translated by 18 translators, using translation time as the dependent variable, and translation modality, target language, their interaction and length of source segment in characters as fixed predictors:

```
edit_time ~ src_len_chr + lang_id * task_type +
(1|subject_id) +
(1 | document_id/item_id) +
(0 + task_type | document_id/item_id)
```

We log-transform the dependent variable, edit time in seconds, given its long right tail. The models are built by adding one element at a time and checking whether such an addition leads to a significantly better model, as indicated by a reduction in AIC (i.e., a decrease of at least 2). Our random effects structure includes random intercepts for different segments (nested with documents) and translators, as well as a random slope for modality over individual segments. We start with an initial model that includes only the two random intercepts (by-translator and by-segment) and proceed by (i) finding significance for nested document/segment random effect; (ii) adding fixed predictors one by one; (iii) adding interactions between fixed predictors; and (iv) adding the random slopes.<sup>9</sup>

Table 8.6 presents the set of predictors included in the final model, along with an estimate of their impact on edit times and their corresponding significance. We find that both PE modalities significantly reduce translation times ( $p < 0.001$ ), with PE<sub>1</sub> being

---

<sup>9</sup>The document processing order was originally included to identify possible longitudinal effects but was removed due to a lack of significant improvements.

Predictor	Estim.	p-value	Sig.
(intercept)	4.92	1.12e-11	***
source length	0.38	< 2e-16	***
lang_ara	-0.49	0.1209	
lang_ita	-0.14	0.6407	
lang_nld	-0.58	0.0733	x
lang_tur	-0.82	0.0162	*
lang_vie	-0.24	0.4254	
task_pe1	-0.49	< 2e-16	***
task_pe2	-0.22	1.77e-07	***
lang_ara:task_pe1	-0.11	0.0505	x
lang_ita:task_pe1	-0.40	8.97e-12	***
lang_nld:task_pe1	-0.41	5.74e-12	***
lang_tur:task_pe1	-0.14	0.0194	*
lang_vie:task_pe1	0.13	0.0290	*
lang_ara:task_pe2	0.05	0.3535	
lang_ita:task_pe2	-0.39	3.30e-11	***
lang_nld:task_pe2	-0.29	4.46e-07	***
lang_tur:task_pe2	0.03	0.5811	
lang_vie:task_pe2	0.04	0.5289	

**Table 8.6 |** LMER modeling results using translation time as the dependent variable. The reference levels for predictors `lang` and `task` are Ukrainian and Translation from scratch (HT), respectively. Estimate impact on edit time for every predictor is provided in log seconds. Significance: \*\*\* = < 0.001, \* = < 0.05, x = < 0.1

significantly faster than PE<sub>2</sub> ( $p < 0.001$ ) across all languages. Considering Ukrainian—the language for which HT is slowest—as the reference level, the reduction in time brought by Google is significantly more pronounced for Italian, Dutch ( $p < 0.001$ ), and Turkish ( $p < 0.05$ ). For mBART-50, however, we only observe significantly more pronounced increases in productivity for Italian and Dutch ( $p < 0.001$ ) compared to the reference. We find these results to corroborate the observations of the previous section.

#### 8.4.2 Post-Editing Rate

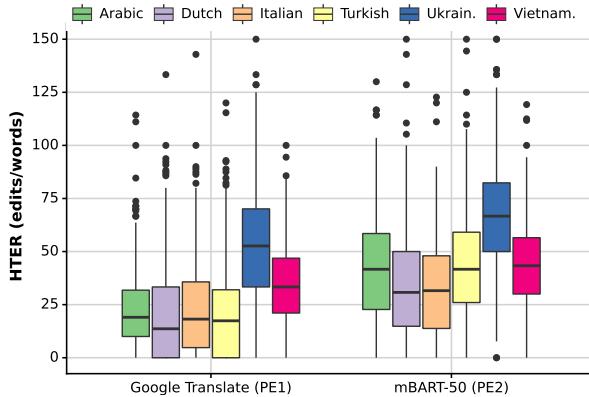
We proceed to study the post-editing patterns using the widely adopted Human-targeted Translation Edit Rate (HTER, Snover et al. (2006)), which is computed as the length-normalized sum of word-level substitutions, insertions, deletions, and shift

operations performed during post-editing.

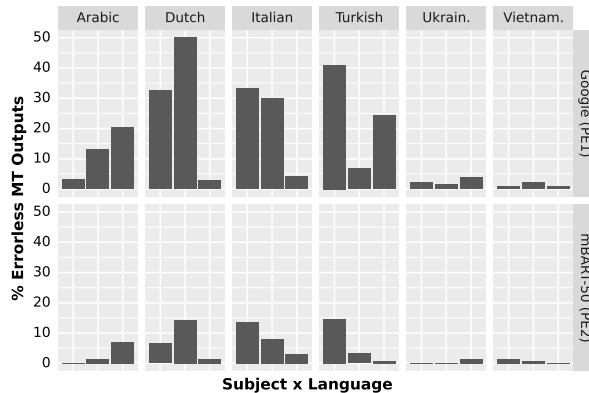
As shown in Figure 8.3, PE<sub>1</sub> required less editing than PE<sub>2</sub> for all languages, and a high variability is observed across the two systems and all languages. Because translators were not informed about the presence of two MT systems, we exclude the possibility that these results reflect an over-reliance or distrust towards a specific MT system. For Google Translate, Ukrainian shows the heaviest edit rate, followed by Vietnamese, whereas Arabic, Dutch, Italian and Turkish all show relatively low amounts of edits. Focusing again on mBART-50 for a more fair cross-lingual comparison, Ukrainian is by far the most heavily edited language, followed by a medium-tier group composed of Vietnamese, Arabic and Turkish, and finally by Dutch and Italian as low-edit languages. Results show that several of our observations on linguistic relatedness and morphology type also apply to edit rates, with languages less related to English or having richer morphology requiring more post-edits on average.

Figure 8.4 visualizes the large gap in edit rates across languages and subjects by presenting the amount of ‘errorless’ MT sentences that were accepted directly, i.e. without any post-editing. We note again how the NMT system significantly influences the rate of occurrence of such sentences, yet nonetheless shows that Dutch and Italian generally present more error-free sentences than Ukrainian and Vietnamese. In particular, for Google Translate outputs, the average rate of error-free sentences is roughly 25% for the former target languages, while for the latter, it accounts for only 3% of total translations. Surprisingly, the English–Turkish pair also fares well, despite the low relatedness between the source and target languages. We note that post-editing effort appears to correlate poorly with the automatic MT quality metrics reported in Table 8.3 (e.g., see the high scores of Vietnamese and the low scores of Dutch PE<sub>1</sub>), highlighting a difficulty in predicting the benefits of MT post-editing over HT for new language pairs.

While HTER is a standard metric adopted in both academic and industrial settings, we also evaluated its character-level variant, CharacTER (Wang et al., 2016), to assess whether it could better account for the editing process of morphologically rich languages. Figure 8.5 presents the CharacTER results. When comparing this plot to the HTER one (Figure 8.3), we notice that CharacTER preserves the overall trends but slightly improves the edit rate for Arabic and Turkish compared to other languages. Nevertheless, we find that HTER correlates slightly better with productivity scores across all tested languages, both at the sentence and document levels.



**Figure 8.3 |** Human-targeted Translation Edit Rate (HTER) for Google Translate and mBART-50 post-editing across available languages.

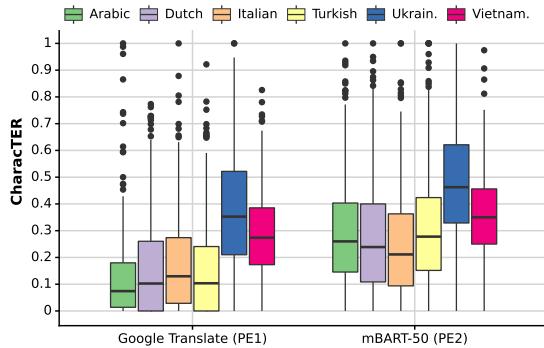


**Figure 8.4 |** Distribution of error-less machine translation sentence outputs (no edits performed during post-editing) for each translator and every language.

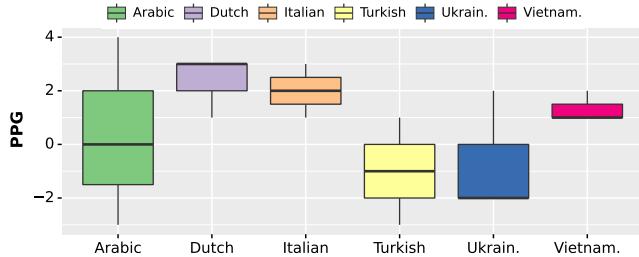
#### 8.4.3 Perception of Productivity Gain

We conclude our analysis by examining the post-task questionnaires, in which participants expressed their perceptions of MT quality and translation speed across HT and PE modalities ( $HT_s$ ,  $PE_s$ )<sup>10</sup> using a 1–7 Likert scale (where 1 is the slowest and 7 is the fastest). We use these to compute the Perceived Productivity Gain (PPG) as  $PPG = PE_s - HT_s$  and visualize it in Figure 8.6. We observe that Italian and Dutch,

<sup>10</sup>We reemphasize that subjects were unaware of the presence of two distinct MT systems.



**Figure 8.5 |** Character-level Human-targeted Translation Edit Rate (CharacTER) for Google Translate and mBART-50 post-editing across available languages.



**Figure 8.6 |** Perceived productivity gains (PPG) between the HT and PE translation modalities, assessed for all subjects after task completion.

the only target languages with marked productivity gains ( $\Delta\text{HT}$ ) regardless of the PE system in Table 8.5, are also the only ones having consistently high ( $\geq 2$ ) PPG scores across all subjects. Moreover, we remark how PPG for target languages with a wide gap in  $\Delta\text{HT}$  scores between high-PE<sub>1</sub> and low-PE<sub>2</sub> (Arabic, Ukrainian) are hardly distinguishable from those of languages in which  $\Delta\text{HT}$  is low for both PE systems (Turkish, Vietnamese). Notably, 4 out of 18 subjects attribute negative PPGs to the PE modality, despite productivity gains being reported across all subjects and languages. These results suggest that worst-case usage scenarios may play an important role in driving PPG, i.e. that *subjects' perception of quality is shaped mainly by particularly challenging or unsatisfying interactions with the NMT system, rather than the average case*. Finally, from the post-task questionnaire, PPG scores exhibit a strong positive correlation with the perception of MT adequacy ( $\rho=0.66$ ), fluency ( $\rho=0.46$ ) and overall quality ( $\rho=0.69$ ), and more generally with a higher enjoyability of PE ( $\rho=0.60$ ), while being inversely correlated with the perception of problematic mistranslations ( $\rho=-0.60$ ).

## 8.5 Limitations

The subjective component introduced by the presence of multiple translators is an important confounding factor in our setup, particularly given the relatively small number of subjects for each language. In our study, we aimed to strike a balance between thorough control of other noise components and faithful reproduction of a realistic translation scenario. However, we recognize that the combination of the limited document context provided by FLORES-101, the variety of topics covered in the texts, and the experimental nature of the PET platform constitutes an atypical setting that may have impacted the translators' natural productivity. Moreover, variability in the content of mBART-50 fine-tuning data, despite their comparable sizes, may have played a role in the observed variability in automatic MT evaluation and PE gains across languages.

## 8.6 Conclusions

We introduced DivEMT, the outcome of a post-editing study that spanned two state-of-the-art NMT systems, involved 18 professional translators, and employed six typologically diverse target languages under a unified setup. We leveraged DivEMT's behavioral data to perform a controlled cross-language analysis of NMT post-editing effort along its temporal and editing effort dimensions. The analysis reveals that NMT drives significant improvements in productivity across all evaluated languages; however, the magnitude of these improvements depends heavily on the language and the underlying NMT system. In this setting, productivity measurements across modalities were found to be generally consistent with the recorded editing patterns. Our results indicate that translators working on language pairs with significant post-editing productivity gains, on average, perform fewer edits and accept more machine-generated translations without any editing. We have also observed a disconnect between post-editing productivity gains and MT quality metrics collected for the same NMT systems. Finally, low source-language relatedness and target morphological complexity seem to hinder productivity when NMT is adopted, even in settings where system architecture and training data are controlled for.

In our qualitative analysis, translators' perception of post-editing usefulness was found to be strongly shaped by problematic mistranslations. Languages showing large productivity gains for both NMT systems were the only ones associated with a positive perception of PE-mediated gains, as opposed to mixed or negative opinions for other translation directions.

Overall, our findings reveal significant variation in post-editing effectiveness across languages and systems, highlighting the need for fine-grained quality assessment tools. In the next chapter, we build upon these insights by conducting a second study with professional post-editors, assessing the impact of word-level error detection methods—including unsupervised approaches that leverage model internals—on the quality and productivity of human post-editing.



# CHAPTER 9

---

## Word-level Quality Estimation for Machine Translation Post-editing

### Chapter Summary

Building upon the user study of the previous chapter, we investigate the impact of word-level quality estimation methods, including the use of model predictive probabilities for uncertainty estimation, on machine translation post-editing. This chapter presents our QE4PE study, involving 42 professional post-editors across two translation directions and four error-span highlight modalities for identifying potential errors in the outputs of a state-of-the-art neural MT model. Post-editing effort and productivity are measured through behavioral logging techniques, while quality improvements are assessed through systematic human annotation. Results reveal that domain, language, and editors' speed are critical factors in determining the effectiveness of highlights, with modest differences between human-made and automated highlights underscoring a gap between accuracy and usability in professional workflows.

This chapter is adapted from the paper *QE4PE: Word-level Quality Estimation for Human Post-Editing* (Sarti et al., 2025b).

*That's just what translation is, I think. That's all speaking is. Listening to the other and trying to see past your own biases to glimpse what they're trying to say. Showing yourself to the world, and hoping someone else understands.*

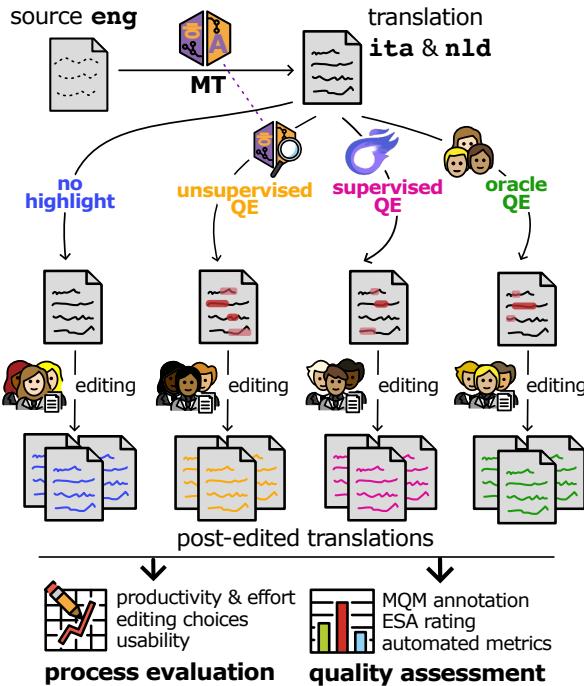
– Rebecca F. Kuang, *Babel* (2022)

## 9.1 Introduction

Recent years saw a steady increase in the quality of machine translation systems and their widespread adoption in professional translation workflows (Kocmi et al., 2024a). Still, human post-editing of MT outputs remains a fundamental step to ensure high-quality translations, particularly for challenging textual domains requiring native fluency and specialized terminology (Liu et al., 2024). Quality estimation (QE) techniques were introduced to reduce post-editing effort by automatically identifying problematic MT outputs without the need for human-written reference translations and were quickly integrated into industry platforms (Tamchyna, 2021).

*Segment-level* QE models correlate well with human perception of quality (Freitag et al., 2024) and exceed the performance of reference-based metrics in specific settings (Rei et al., 2021; Amrhein et al., 2022, 2023). On the other hand, *word-level* QE methods for identifying error spans requiring revision have received less attention in the past due to their modest agreement with human annotations, despite their promise for more granular and interpretable quality assessment in line with modern MT practices (Zerva et al., 2024). In particular, while the accuracy of these approaches is regularly assessed in evaluation campaigns, research has rarely focused on assessing the impact of such techniques in realistic post-editing workflows, with notable exceptions suggesting limited benefits (Shenoy et al., 2021; Eo et al., 2022). This hinders current QE evaluation practices: by foregoing experimental evaluation with human editors, it is implicitly assumed that word-level QE will become helpful once sufficient accuracy is achieved, without accounting for the additional challenges associated with successfully integrating these methods into post-editing workflows.

In this chapter, we present a study, which we dub QE4PE (Quality Estimation for Post Editing), addressing this gap through a large-scale study with 42 professional translators for the English→Italian and English→Dutch directions to measure the impact of word-level QE on editing quality, productivity and usability. We aim for a realistic and reproducible setup, employing the high-quality open-source NLLB 3.3B MT model (NLLB Team et al., 2024) to translate challenging documents from biomedical and social media domains. We then conduct a controlled evaluation of post-editing with error spans in four *highlight modalities*, i.e. using highlights derived from four word-level QE methods: a **supervised** state-of-the-art QE model trained on human error annotations (**xCOMET**, Guerreiro et al., 2024), an **unsupervised** method leveraging the uncertainty of the MT model during generation, **oracle** error spans obtained from the consensus of previous human post-editors, and a **no highlight** baseline. The human post-editing is performed using GROTE, a simple online interface we built to support the real-time



**Figure 9.1 |** A summary of the QE4PE study. Documents are translated by a neural MT model and reviewed by professional editors across two translation directions and four highlight modalities. Editing effort, productivity and usability across modalities are estimated from editing logs and questionnaires. Finally, the quality of MT and edited outputs is assessed with MQM/ESA human annotations and automatic metrics.

logging of granular editing data, enabling a quantitative assessment of editing effort and productivity across highlight modalities. We also survey professionals using an online questionnaire to collect qualitative feedback about the usability and quality of the MT model, as well as the interface and error span highlights. Finally, a subset of the original MT outputs and their post-edited variants is annotated following the MQM and ESA protocols (Lommel et al., 2013; Kocmi et al., 2024b) to verify quality improvements after post-editing. See Figure 9.1 for an overview of the study. Most similar to our study, Shenoy et al. (2021) investigated the effect of synthetic word-level QE highlights for English→German post-editing on Wikipedia data, concluding that word-level QE accuracy was at the time still insufficient to produce tangible productivity benefits in human editing workflows. We expand the scope of such evaluation by including two translation directions, two challenging real-world text domains and state-of-the-art

MT and QE systems and methods. In summary, our work represents a step towards evaluating translation technologies and methods that leverage model internals, centered on users' experience (Guerberof-Arenas and Moorkens, 2023; Savoldi et al., 2025).

We release all data, code and the GROTE editing interface to foster future studies on the usability of error span highlighting techniques for other word-level QE methods and translation directions.<sup>1</sup>

## 9.2 Experimental Setup

### 9.2.1 Structure of the Study

Our study is organized in five stages:

**1) Oracle Post-editing** As a preliminary step, segments later used in the main assessment are post-edited by three professionals per direction using their preferred interface without logging. This allows us to obtain post-edits and produce **oracle** word-level spans based on the editing consensus of multiple human professionals. Translators involved in this stage are not involved further in the study.

**2) Pretask (Pre)** The pretask allows the **core translators** (12 per language direction, see Section 9.2.4) to familiarize themselves with the GROTE interface and text highlights. Before starting, all translators complete a questionnaire to provide demographic and professional information about their profile (Table 9.9). In the pretask, all translators work in an identical setup, post-editing a small set of documents similar to those of the main task with **Oracle** highlights. We assign core translators into three groups based on their speed, as determined by their editing logs (4 translators per group for *faster*, *average* and *slower* groups in each direction). Individuals from each group are then randomly assigned to each highlight modality to ensure an equal representation of editing speeds, resulting in 1 *faster*, 1 *average*, and 1 *slower* translator for each highlight modality. This procedure is repeated independently for both translation directions.

**3) Main Task (Main)** This task, conducted in the two weeks following the pretask, covers the majority of the collected data and is the main object of study for the analyses of Section 9.3. In the main task, 24 core translators work on the same texts using the GROTE interface, with three translators per modality in each translation direction, as shown in Figure 9.1. After the main task, translators complete a questionnaire on the

---

<sup>1</sup>Dataset: [gsarti/qe4pe](https://github.com/gsarti/qe4pe); Interface: [gsarti/grote](https://github.com/gsarti/grote); Code: <https://github.com/gsarti/qe4pe>

quality and usability of the MT outputs, the interface and, where applicable, word highlights.<sup>2</sup>

**4) Post-Task (Post)** After MAIN, the 12 core translators per direction are asked to post-edit an additional small set of related documents with GROTE, but this time working all with the **No Highlight** modality. This step lets us obtain baseline editing patterns for each translator, allowing us to estimate individual speed and editing differences across highlight modalities without the confounding effect of interface proficiency accounted for in the PRE stage.

**5) Quality Assessment (QA)** Finally, a subset consisting of 148 main task segments is randomly selected for manual annotation by six new translators per direction (see Section 9.2.4). For each segment, the original MT output and all its post-edited versions are annotated with MQM error spans, including minor/major error severity and a subset of MQM error categories, including e.g., mistranslations, omissions, and stylistic errors (Lommel et al., 2013). Table 9.1 presents our annotation guidelines.<sup>3</sup> Moreover, the annotator proposes corrections for each error span, ultimately providing a 0-100 quality score, which matches the common DA scoring adopted in multiple WMT campaigns. We adopt this scoring system, which closely adheres to the ESA evaluation protocol (Kocmi et al., 2024b), following recent results showing its effectiveness and efficiency for ranking MT systems.

In summary, for each translation direction, we collect three full sets of oracle post-edits, 12 full sets of edits with behavioral logs for PRE, MAIN and POST task data, and 13 subsets of main task data (12 post-edits, plus the original MT output) annotated with MQM error spans, corrections and segment-level ESA ratings. Moreover, we also collect 12 pre- and post-task questionnaire responses from *core set* translators to obtain a qualitative view of the editing process.

### 9.2.2 Highlight Modalities

We conduct our study on four highlight modalities across two severity levels (*minor* and *major* errors). Using multiple severity levels aligns with current MT evaluation practices (Freitag et al., 2021a, 2024), as well as previous results indicating that users tend to prefer more granular and informative word-level highlights (Shenoy et al., 2021; Vasconcelos et al., 2025). The highlight modalities we employ are:

---

<sup>2</sup>We do not disclose the highlight modality to translators to avoid biasing their judgment in the evaluation.

<sup>3</sup>See Table C.15 for annotation examples.

Error category	Subcategory	Description
<b>Accuracy</b> Incorrect meaning has been transferred to the source text.	<b>Addition</b>	Translation includes the information that is not present in the source and it changes or distorts the original message.
	<b>Omission</b>	Translation is missing the information that is present in the source, which is important to convey the message.
	<b>Mistranslation</b>	Translation does not accurately represent the source content meaning.
	<b>Inconsistency</b>	There are internal inconsistencies in the translation (for example, using different verb forms in the bullet list or in CTAs, calling the same UI element differently, terminology used inconsistently etc).
<b>Linguistic</b> Official linguistic reference sources such as grammar books.	<b>Untranslated</b>	Content that should have been translated has been left untranslated.
	<b>Punctuation</b>	Punctuation is used incorrectly (for the locale or style), including missing or extra white spaces and the incorrect use of space (non-breaking space). Violation of typographic conventions of the locale.
	<b>Spelling</b>	Issues related to spelling of words, including typos, wrong word hyphenation, word breaks and capitalization.
<b>Style</b> Not suitable/native; too literal or awkward.	<b>Grammar</b>	Issues related to the grammar or syntax of the text, other than spelling.
	<b>Inconsistent Style</b>	Style is inconsistent within a text.
	<b>Readability</b>	Translation does not read well (due to heavy sentence structure, frequent repetitions, unidiomatic).
<b>Wrong Register</b>		Inappropriate style for the specific subject field, the level of formality, and the mode of discourse (e.g., written text versus transcribed speech).
Severity level	Description	
<b>Major</b>	The Severity Level of an error that seriously affects the understandability, reliability, or usability of the content for its intended purpose or hinders the proper use of the product or service due to a significant loss or change in meaning or because the error appears in a highly visible or important part of the content.	
<b>Minor</b>	The Severity Level of an error that does not seriously impede the usability, understandability, or reliability of the content for its intended purpose, but has a limited impact on, for example, accuracy, stylistic quality, consistency, fluency, clarity, or general appeal of the content.	
<b>Neutral</b>	The Severity Level of an error that differs from a quality evaluator's preferential translation or that is flagged for the translator's attention but is an acceptable translation.	

Table 9.1 | Annotation instructions for our MQM-inspired error taxonomy.

**No Highlight** The text is presented as-is, without any highlighted spans. This setting serves as a baseline to estimate the default post-editing quality and productivity using our interface.

**Oracle** Following the Oracle Post-editing phase, we produce oracle error spans from the editing consensus of human post-editors. We label text spans that were edited by two of the three translators as *minor*, and those edited by all three translators as *major*, following the intuition that more critical errors are more likely to be identified by several annotators, while minor changes will show more variance across subjects. This modality serves as a best-case scenario, providing an upper bound for future improvements in word-level QE quality.

**Supervised** In this setting, word-level error spans are obtained using xCOMET-XXL (Guerreiro et al., 2024), which is a multilingual transformer encoder (Goyal et al., 2021) further trained for joint word- and sentence-level QE prediction. We select xCOMET-XXL in light of its broad adoption, open accessibility and state-of-the-art performance in QE across several translation directions (Zerva et al., 2024). For the severity levels, we use the labels predicted by the model, mapping *critical* labels to the *major* level.

**Unsupervised** In this modality, we exploit the access to the MT model producing the original translations to obtain *uncertainty-based highlights*. As a preliminary evaluation to select a capable unsupervised word-level QE method, we evaluate two unsupervised QE methods employing token log-probabilities assigned by MT model to predict human post-edits: raw negative log-probabilities (LOGPROBS), corresponding to the surprisal assigned by the MT model to every generated token, and their variance for 10 steps of Monte Carlo Dropout (LOGPROBS<sub>mcd var</sub>, Gal and Ghahramani, 2016). We employ surprisal-based metrics, following previous work that shows their effectiveness in predicting translation errors (Fomicheva and Specia, 2019) and human editing time (Lim et al., 2024). We collect scores for the English→Italian and English→Dutch directions of QE4PE Oracle post-edits and our DivEMT dataset (Sarti et al., 2022) to identify the best-performing method, using metric scores extracted from the original models used for translation to predict human post-edits. We use average precision (AP) as a threshold-agnostic performance metric for the tested continuous methods. Oracle highlights obtained from the consensus of three annotators in the first stage of the study are used as reference for QE4PE, while a single set of post-edits is available for DivEMT. The xCOMET-XXL model used for Supervised highlights, and the average agreement of individual Oracle editors with the consensus label are also included for comparison.

Method	DivEMT				QE4PE			
	En→It		En→Nl		En→It		En→Nl	
	AP	AU	AP	AU	AP	AU	AP	AU
LOGPROBS (Fomicheva et al., 2020)	0.18	0.18	0.19	0.19	0.10	0.09	0.09	0.09
LOGPROBS <sub>MCD VAR</sub> (Fomicheva et al., 2020, <b>Unsup.</b> )	0.41	0.41	0.42	0.42	0.23	0.23	0.31	0.31
XCOMET-XXL (Guerreiro et al., 2024, <b>Sup.</b> )	-	-	-	-	0.16	0.23	0.19	0.28
AVG. <b>Oracle</b> SINGLE TRANSLATOR	-	-	-	-	0.53	0.73	0.55	0.75

**Table 9.2 |** Average Precision (AP) and Area Under the Precision-Recall Curve (AU) between metrics and error spans derived from human post-editing. We use mBART 1-to-50 and NLLB 3B respectively for DivEMT and QE4PE. For DivEMT, a single post-editor is available for computing the agreement, while for QE4PE we use consensus-based **ORACLE** highlights. For QE4PE, we report the average agreement between individual oracle post-editors and their consensus as an agreement upper bound.

Results in Table 9.2 show a strong performance for the LOGPROBS<sub>mcd var</sub> method, even surpassing the accuracy of the supervised XCOMET model across both datasets.<sup>4</sup> Hence, we select it for the **Unsupervised** highlight modality, setting value thresholds for minor-/major errors to match the respective highlighted word proportions in the **Supervised** modality to ensure a fair comparison of their effectiveness in the post-editing task.

### 9.2.3 Data and MT model

**MT Model** On the one hand, the MT model must achieve *high translation quality* in the selected languages to ensure our experimental setup applies to state-of-the-art proprietary systems. Still, the MT model should be *open-source* and have a *manageable size* to ensure reproducible findings and enable the computation of uncertainty for the unsupervised setting. All considered, we use NLLB 3.3B (NLLB Team et al., 2024), a widely used MT model achieving industry-level performances across 200 languages (Moslem et al., 2023).

**Data selection** We begin by selecting two translation directions, English→Italian and English→Dutch, according to the availability of professional translators from our industrial partners. We intentionally focus on out-of-English translations as they are generally more challenging for modern MT models (Kocmi et al., 2023). We aim to identify documents that are manageable for professional translators without domain-specific expertise but still prove challenging for our MT model to ensure a sufficient

<sup>4</sup>Highlights are extended from tokens to words to match the granularity of other modalities.

Task	Domain	# Docs	# Seg.	# Words
PRE	Social	4	23	539
	Biomed.	2	15	348
MAIN	Social	30	160	3375
	Biomed.	21	165	3384
POST	Social	6	34	841
	Biomed.	2	16	257
<b>Total</b>		64	413	8744

Table 9.3 | Statistics for QE4PE data.

amount of error spans across modalities. Since original references for our selected translation direction were not available, we do not have a direct mean to compare MT quality in the two languages. However, according to our human MQM assessment in Section 9.3.3 (Table 9.7), NLLB produces a comparable amount of errors across Dutch and Italian translations, suggesting similar quality.

We begin by translating 3,672 multi-segment English documents from the WMT23 General and Biomedical MT shared tasks (Kocmi et al., 2023; Neves et al., 2023) and MT test suites to Dutch and Italian. Our choice for these specialized domains, as opposed to e.g. generic news articles, is driven by the real-world needs of the translation industry for domain-specific post-editing support (Eschbach-Dymanus et al., 2024; Li et al., 2025). Moreover, focusing on domains that are considerably more challenging for MT systems than news, as shown by recent WMT campaigns (Neves et al., 2024), ensures a sufficient amount of MT errors to support a sound comparison of word-level QE methods. Then, `XCOMET-XXL` is used to produce a first set of segment-level QE scores and word-level error spans for all segments. To make the study tractable, we further narrow down the selection of documents according to several heuristics to ensure a realistic editing experience and a balanced occurrence of error spans (details in Section C.2.1). This procedure yields 351 documents, from which we manually select a subset of 64 documents (413 segments, 8,744 source words per post-editor) across two domains:

- **Social media posts**, including Mastodon posts from the WMT23 General Task (Kocmi et al., 2023) English↔German evaluation and Reddit comments from the Robustness Challenge Set for Machine Translation (RoCS-MT, Bawden and Sagot, 2023), displaying atypical language use, such as slang or acronymization.
- **Biomedical abstracts** extracted from PubMed from the WMT23 Biomedical

Source <sub>EN</sub>	So why is it that people jump through extra hoops to install Google Maps?
<b>No High.</b>	Quindi perché le persone devono fare un salto in più per installare Google Maps?
<b>Oracle</b>	Quindi perché le persone <b>devono fare un salto in più</b> per installare Google Maps?
<b>Sup.</b>	Quindi perché le persone devono <b>fare un salto in più</b> per installare Google Maps?
<b>Unsup.</b>	Quindi perché le persone <b>devono fare un salto in più</b> per installare Google Maps?
PE <sub>No High.</sub>	Quindi perché le persone devono fare un <b>passaggio</b> in più per installare Google Maps?
PE <sub>Oracle</sub>	Allora, perché le persone <b>fanno un passaggio</b> in più per installare Google Maps?
PE <sub>Sup.</sub>	Quindi perché le persone <b>fanno passaggi</b> in più per installare Google Maps?
PE <sub>Unsup.</sub>	Quindi perché le persone <b>fanno i salti mortali</b> per installare Google Maps?

**Table 9.4** |  $\text{EN} \rightarrow \text{IT}$  example from the QE4PE dataset, showing **minor/major** word highlights and a single post-edit per modality, with modified words **highlighted**.

Remove negation (13-6)	
English	<b>No significant differences</b> were found with [...]
Dutch	Er werden <b>geen significante verschillen</b> → <b>significante verschillen</b> gevonden met [...]
Title literal translation (16-3)	
English	The <b>Last of Us</b> is an easy and canonical example of dad-ification. [...]
Italian	The <b>Last of Us</b> → L'ultimo di noi è un esempio facile e canonico di dad-ification. [...]
Wrong term (48-5)	
English	[...], except for <b>alkaline phosphatase</b> .
Italian	[...], ad eccezione della <b>fosfatasi alcalina</b> → <b>chinasi proteica</b> .

**Table 9.5** | Examples of original → manually inserted critical errors with document-segment ID from Table 9.8.

Translation Task (Neves et al., 2023), including domain-specific terminology.

Table 9.3 present statistics for the PRE, MAIN and Post editing stages, and Table 9.4 shows an example of highlights and edits. While including multiple domains in the same task can render our post-editing setup less realistic, we deem it essential to test the cross-domain validity of our findings.

**Critical Errors** Before producing highlights, we manually introduce 13 critical errors into main task segments to assess post-editing thoroughness. Errors are produced, for example, by negating statements, inverting the polarity of adjectives, inverting numbers, and corrupting acronyms. We replicate the errors in both translation directions to enable direct comparison. Most of these errors were correctly identified across all three highlight modalities. Table 9.5 presents some examples of critical errors in the dataset.

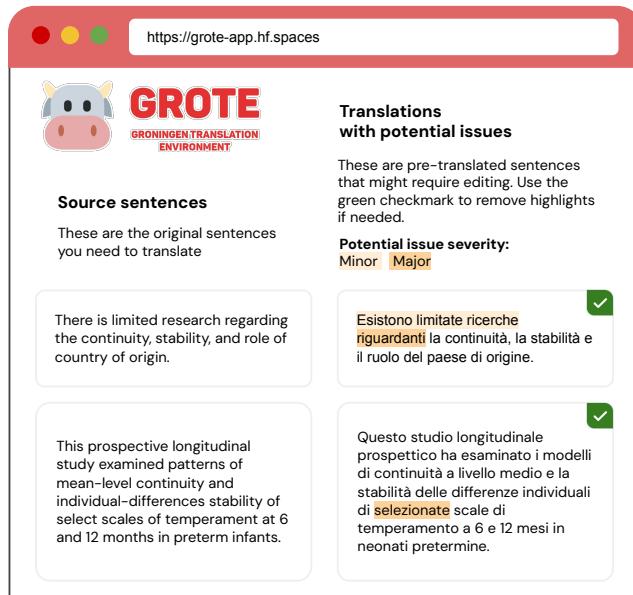
### 9.2.4 Participants

For both directions, the professional translation companies Translated and Global Textware recruited three translators for the Oracle post-editing stage, the core set of 12 translators working on PRE, MAIN and POST tasks, and six more translators for the QA stage, for a total of 21 translators per direction. All translators were freelancers with native proficiency in their target language and self-assessed proficiency of at least C1 in English. Almost all translators had more than two years of professional translation experience and regularly post-edited MT outputs (details in Table 9.9).

### 9.2.5 Editing Interface

We develop a custom interface, which we name **Groningen Translation Environment** (GroTE, Figure 9.2), to support editing over texts with word-level highlights. While the MMPE tool used by Shenoy et al. (2021) provide extensive multimodal functionalities (Herbig et al., 2020), we aim for a bare-bones setup to avoid confounders in the evaluation. GroTE is a web interface based on Gradio (Abid et al., 2019) and hosted on the Hugging Face Spaces to enable multi-user data collection online. Upon loading a document, source texts and MT outputs for all segments are presented in two columns, following standard industry practices. For modalities with highlights, the interface provides an informative message and supports removing all highlights from a segment via a button, with highlights on words disappearing automatically upon editing, as in (Shenoy et al., 2021). The interface supports real-time logging of user actions, allowing for the analysis of the editing process. In particular, we log the start and end times for each edited document, the accessing and exiting of segment textboxes, highlight removals, and individual keystrokes during editing.

GroTE intentionally lacks standard features such as translation memories, glossaries, and spellchecking to ensure equal familiarity among translators, ultimately controlling for editor proficiency with these tools, as done in previous studies (Shenoy et al., 2021; Sarti et al., 2022). While most translators noted the lack of advanced features in our usability assessment, the majority also found the interface easy to set up, access, and use (Table 9.9).



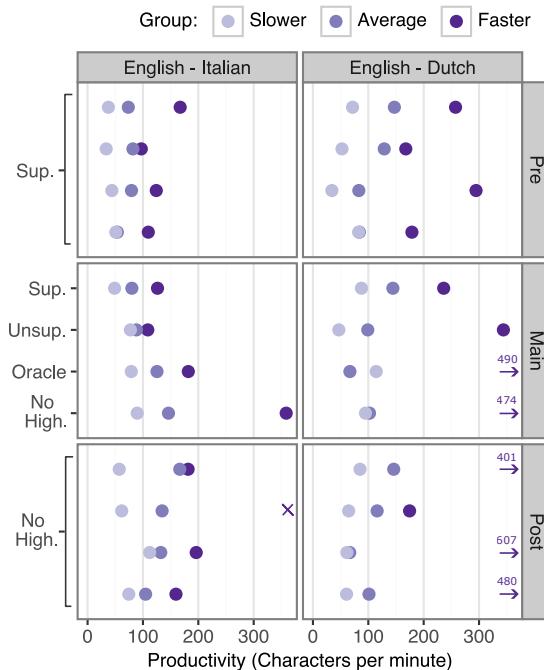
**Figure 9.2 |** An example of the QE4PE GROTE setup for two segments in an English→Italian document.

## 9.3 Analysis

### 9.3.1 Productivity

We obtain segment- and document-level edit times and compute editing *productivity* as the number of processed source characters divided by the sum of all document-level edit times, measured in characters per minute. To account for potential breaks taken by post-editors during editing, we filter out pauses between logged actions longer than 5 minutes. We note that this procedure does not significantly impact the overall ranking of translators, while ensuring a more robust evaluation of editing time.

**Do Highlights Make Post-editors Faster?** Figure 9.3 shows translators' productivity across stages, with every dot corresponding to the productivity of a single individual. We observe that no highlight modality consistently leads to faster editing across all speed groups and that the ordering of PRE-task speed groups is maintained in the following stages despite the different highlight modalities. These results suggest that individual variability in editing speed is more critical than highlight modality in predicting editing speed. However, the fastest English→Dutch translators achieved outstanding



**Figure 9.3 |** Productivity of post-editors across QE4PE stages (PRE, MAIN, POST). The → marks entries with speed > 300 characters per minute and X marks missing data. Each row corresponds to the same three translators across all stages.

productivity, i.e. > 2 standard deviations above the overall mean (entries with > 300 char/min, → in Figure 9.3) almost exclusively in **No Highlight**, and, **Oracle** modalities, suggesting that lower-quality highlights hinder editing speed.

We validate these observations by fitting a negative binomial mixed-effect model on segment-level editing times (model details in Table C.10). Excluding random factors such as translator and segment identity from the model results in a significant drop in explained variance, confirming the inherent variability of editing times ( $R^2 = 0.93 \rightarrow 0.41$ ). Model coefficients indicate that the output length of MT and the proportion of highlighted characters are the primary factors driving an increase in editing times, possibly reflecting an increase in cognitive effort required to process additional information. We find that highlights have a significant impact on increasing the editing speed of English→Italian translators ( $p < 0.001$ ), but a minimal impact for English→Dutch. Comparing the productivity of the same translator editing with and without highlights (MAIN vs POST), two-thirds of the translators editing with highlights

were up to two times slower on biomedical texts. However, the same proportion of translators was up to three times faster on social media texts across both directions.

In summary, we find that **highlight modalities are not predictive of edit times on their own**, but translation direction and domain play an important role in determining the effect of highlights on editing productivity. We attribute these results to two main factors, which will remain central in the analysis of the following sections: (1) the different *propensity of translators to act upon highlighted issues* in the two tested directions, and (2) the different *nature of errors highlighted across domains*.

### 9.3.2 Highlights and Edits

We then examine how highlights are distributed across modalities and how they influence the editing choices of human post-editors.

**Agreement Across Modalities** First, we quantify how different modalities agree in terms of highlights' distribution and editing. We find that highlight overlaps across modalities range from 15% to 39% when comparing highlight modalities in a pairwise fashion, with the highest overlap for English→Italian social media and English→Dutch biomedical texts.<sup>5</sup> Despite the relatively low agreement on highlights, we find an average agreement of 73% for post-edited characters across modalities. This suggests that edits are generally uniform regardless of highlight modalities and are not necessarily restricted to highlighted spans.<sup>6</sup>

**Do Highlights Accurately Identify Potential Issues?** Table 9.6 (Base Freq.) shows raw highlight and edit frequencies across modalities. We observe different trends across the two language pairs: for English→Italian, post-editors working with highlights edit more than twice as much as those working with **No Highlight**, regardless of the highlight modality. On the contrary, for English→Dutch they edit 33% less in the same setting. These results suggest a different attitude towards acting upon highlighted potential issues across the two translation directions, with English→Italian translators appearing to be more inclined to perform more edits when highlights are present. We introduce four metrics to quantify highlights-edits overlap:

- $P(E|H)$  and  $P(H|E)$ , reflecting highlights' *precision* and *recall* in predicting edits, respectively.

---

<sup>5</sup>Scores are normalized to account for highlight frequencies across modalities. Agreement is shown in Table C.12.

<sup>6</sup>Editing agreement is shown in Figure C.6.

	Base Freq.		Measured				Projected			
	$P(H)$	$P(E)$	$P(E H)$	$\Lambda_E$	$P(H E)$	$\Lambda_H$	$\vec{P}(E H)$	$\vec{\Lambda}_E$	$\vec{P}(H E)$	$\vec{\Lambda}_H$
<b>English → Italian</b>										
<b>No High.</b>	-	0.05	-	-	-	-	-	-	-	-
Random	0.16	-	-	-	-	-	0.06	1.20	0.18	1.20
<b>Oracle</b>	0.15	0.12	<b>0.37</b>	<b>4.62</b>	<b>0.45</b>	<b>4.1</b>	<b>0.18</b> <sub>↓0.2</sub>	<b>6.00</b> <sub>↑1.4</sub>	<b>0.55</b> <sub>↑0.1</sub>	<b>4.23</b> <sub>↑0.1</sub>
<b>Unsup.</b>	0.16	0.13	0.25	2.27	0.21	2.2	0.11 <sub>↓0.1</sub>	2.75 <sub>↑0.5</sub>	0.37 <sub>↑0.2</sub>	2.47 <sub>↑0.3</sub>
<b>Sup.</b>	0.12	0.16	0.28	2.00	0.22	2.0	0.14 <sub>↓0.1</sub>	3.50 <sub>↑1.5</sub>	0.35 <sub>↑0.1</sub>	3.18 <sub>↑1.2</sub>
<b>English → Dutch</b>										
<b>No High.</b>	-	0.14	-	-	-	-	-	-	-	-
Random	0.17	-	-	-	-	-	0.16	1.14	0.19	1.19
<b>Oracle</b>	0.20	0.10	<b>0.26</b>	<b>4.33</b>	<b>0.53</b>	3.12	<b>0.28</b> <sub>↑0.0</sub>	<b>2.55</b> <sub>↓1.8</sub>	<b>0.40</b> <sub>↓0.1</sub>	2.35 <sub>↓0.8</sub>
<b>Unsup.</b>	0.20	0.11	0.20	2.50	0.36	2.00	0.22 <sub>↑0.0</sub>	1.83 <sub>↓0.7</sub>	0.31 <sub>↓0.1</sub>	1.72 <sub>↓0.3</sub>
<b>Sup.</b>	0.12	0.09	0.24	3.43	0.33	<b>3.30</b>	<b>0.28</b> <sub>↑0.0</sub>	2.33 <sub>↓1.1</sub>	0.24 <sub>↓0.1</sub>	<b>2.40</b> <sub>↓0.9</sub>

**Table 9.6 |** Highlighting ( $H$ ) and editing ( $E$ ) average statistics across directions and highlight modalities. **Measured:** actual edits performed in the specified modality. **Projected:** using modality highlights over **No HIGHLIGHT** edits to account for editing biases (Section 9.3.2). Random highlights matching average word frequencies are used as random baseline, and Projected increases<sub>↑</sub>/decreases<sub>↓</sub> compared to Measured counterparts are shown. Significant **Oracle** gains over all other modalities are underlined ( $p < 0.05$  with Bonferroni correction).

- $\Lambda_E \stackrel{\text{def}}{=} P(E|H)/P(E|\neg H)$  shows how much more likely an edit is to fall within rather than outside highlighted characters.
- $\Lambda_H \stackrel{\text{def}}{=} P(H|E)/P(H|\neg E)$  shows how much more likely it is for a highlight to mark edited rather than unmodified spans.

Intuitively, character-level recall  $P(H|E)$  should be more indicative of highlight quality compared to precision  $P(E|H)$ , provided that word-level highlights can be useful even when not minimal. For example, if the fully-highlighted word *traduttore* is changed to its feminine version *traduttrice*,  $P(H|E) = 1$  (edit correctly and fully predicted) but  $P(E|H) = 0.3$  since word stem characters are left unchanged. Table 9.6 (Measured) shows metric values across the three highlight modalities<sup>7</sup>. As expected, **Oracle** highlights obtain the best performance in terms of precision and recall, with  $P(H|E)$ , in particular, being significantly higher than the other two modalities across both directions.

<sup>7</sup>Breakdowns by domain and speed shown in Table C.14 and Table C.13

Surprisingly, we find no significant precision and recall differences between **Oracle** and **Unsupervised** highlights, despite the word-level QE training of xCOMET used in the former modality. Moreover, they support the potential of unsupervised, model internals-based techniques to complement or substitute more expensive supervised approaches. Still, likelihood ratios  $\Lambda_E, \Lambda_H \gg 1$  for all modalities and directions indicate that highlights are 2–4 times more likely to precisely and comprehensively encompass edits than non-highlighted texts. This suggests that even imperfect highlights that do not reach **Oracle**-level quality might effectively direct editing efforts toward potential issues. We validate these observations by fitting a zero-inflated negative binomial mixed-effects model to predict segment-level edit rates. Results confirm a significantly higher edit rate for English→Italian highlighted modalities and the social media domain with  $p < 0.001$  (features and significances shown in Appendix Table C.11). We find a significant zero inflation associated with translator identity, suggesting the choice of leaving MT outputs unedited is highly subjective.

**Do Highlights Influence Editing Choices?** Since we found in Section 9.3.1 that the proportion of highlighted characters affects the editing rate of translators, we question whether the relatively high  $P(E|H)$  and  $P(H|E)$  values might be artificially inflated by translators’ eagerness to intervene on highlighted spans. In other words, do highlights identify actual issues, or do they condition translators to edit when they otherwise would not? To answer this, we propose to *project* highlights from a selected modality—in which highlights were shown during editing—onto the edits performed by the **No Highlight** translators on the same segments. The resulting difference between measured and projected metrics can then be used as an estimate of the impact of highlight presentation on their resulting accuracy.

To further ensure the soundness of our analysis, we use a set of projected Random highlights as a lower bound for highlight performance. To make the comparison fair, Random highlights are created by randomly highlighting words in MT outputs to match the average word-level highlight frequency across all highlighted modalities, given the current domain and translation direction. Table 9.6 (Projected) shows results for the three highlighted modalities. First, all projected metrics remain consistently above the Random baseline, suggesting a higher-than-chance ability to identify errors even for worst-performing highlight modalities. Projected precision scores  $P(E|H)$  depend on edit frequency, and hence see a major decrease for English→Italian, where the **No Highlight** edit rate  $P(E)$  is much lower. However, the increase in  $\Lambda_E$  across all English→Italian modalities confirms that, despite the lower edit proportion, highlighted texts remain notably more likely to be edited than non-highlighted ones. Conversely,

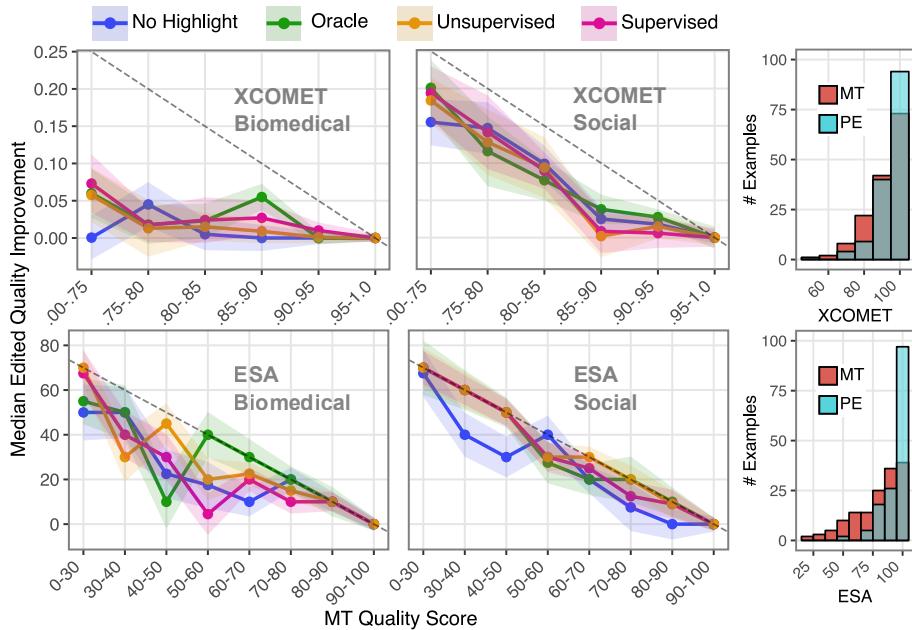
the lower  $\Lambda_E$ ,  $P(H|E)$  and  $\Lambda_H$  for English→Dutch show that edits become much less skewed towards highlighted spans in this direction when accounting for presentation bias.

Overall, while the presence of highlights makes English→Italian translators more likely to intervene in MT outputs, their location in the MT output often pinpoints issues that would be edited regardless of the highlighting. English→Dutch translators, on the contrary, intervene at roughly the same rate regardless of highlight presence, but their edits are focused mainly on highlighted spans when they are present. This difference is consistent across all subjects in both directions, despite the identical setup and comparable MT and QE quality across languages. This suggests that cultural factors may play a non-trivial role in determining the usability and influence of QE methods regardless of span accuracy, a phenomenon previously observed in human–AI interaction studies Ge et al. (2024).

### 9.3.3 Quality Assessment

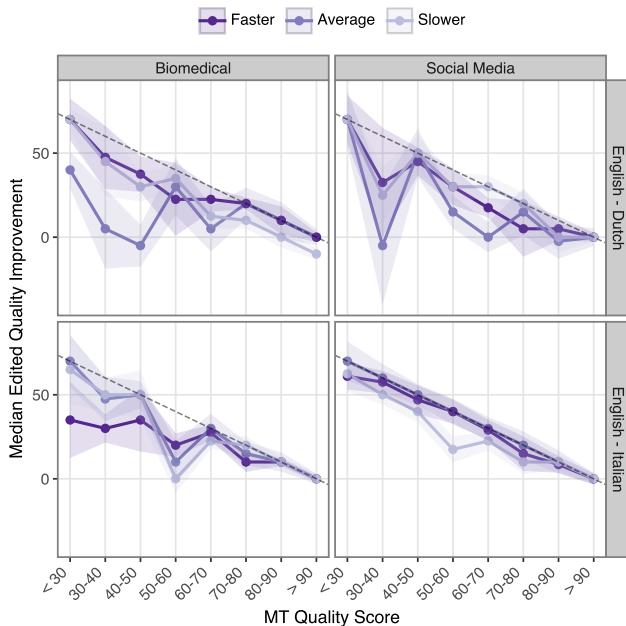
We continue our assessment by inspecting the quality of MT and post-edited outputs along three dimensions. First, we use xCOMET segment-level QE ratings as an automatic approximation of quality and compare them to human-annotated quality scores collected in the last phase of our study. For efficiency, these are obtained for the 0–100 Direct Assessment scale commonly used in QE evaluation (Specia et al., 2020), but following an initial step of MQM error annotation to condition scoring on found errors, as prescribed by the ESA protocol (Kocmi et al., 2024b). Then, MQM error span annotations are used to analyze the distribution of error categories. Finally, we manually assess critical errors, which were inserted to quantify the effect of highlighting modalities on unambiguous issues.

**Do Highlights Influence Post-Editing Quality?** In this stage, we focus particularly on *edited quality improvements*, i.e. how post-editing the same MT outputs under different highlight conditions influences the resulting quality of translations. We operationalize this assessment using human ratings and automatic metrics to score MT and post-edited translations, using their difference as the effective quality gain after the post-editing stage. Scores for this metric are generally positive—that is, human post-editing improves quality—and are bounded by the maximal achievable quality gain given the initial MT quality. Figure 9.4 shows median improvement values across quality bins defined from the distribution of initial MT quality scores (shown in histograms), in which all post-edited versions of each MT output appear as separate observations. Positive median



**Figure 9.4 |** Median quality improvement for post-edited segments at various initial MT quality levels across domains and highlight modalities. Quality scores are estimated using XCOMET segment-level QE (top) and professional ESA annotations (bottom). Histograms show example counts across quality bins for the two metrics. Dotted lines show upper bounds for quality improvements given starting MT quality.

scores confirm that post-edits generally lead to improvements in quality across all tested settings. However, we observe different trends across the two metrics: across both domains, XCOMET greatly underestimates the human-assessed ESA quality improvement, especially for biomedical texts, where it shows negligible improvement regardless of the initial MT quality. These results echo recent findings cautioning users against the poor performance of trained MT metrics for unseen domains and high-quality translations (Agrawal et al., 2024; Zouhar et al., 2024). Focusing on the more reliable ESA scores, we observe large quality improvements from post-editing, as indicated by near-maximal quality gains across most bins and highlight modalities. While **No Highlight** seems to underperform other modalities in the social media domain, the lack of more notable differences in gains across highlight modalities suggests that **highlights' quality impact might not be evident in terms of segment-level quality**, motivating our next steps in the quality analysis.



**Figure 9.5 |** Median ESA quality improvement following post-editing for segments at various initial MT quality levels across translators’ speed groups, showing no clear quality trends across editors’ productivity levels.

We also find no clear relationship between translator speed and improved edited quality, suggesting that higher productivity does not come at a cost for faster translators (Figure 9.5). This finding confirms that neglecting errors is not the cause of the different editing patterns observed in previous sections.

**Which Error Types Do Highlights Identify?** Table 9.7 shows a breakdown of MQM annotations for MT and all highlight modalities using the *Accuracy*, *Style* and *Linguistic* macro-categories of MQM errors.<sup>8</sup> At this granularity, differences across modalities become visible, with overall error counts showing a clear relation to  $\Delta_E$  from Table 9.6, i.e. **Oracle** being remarkably better for English→Italian, with milder and more uniform trends in English→Dutch. At least for English→Italian, these results confirm that an observable quality improvement from editing with highlights is present in the best-case **Oracle** scenario. By contrast, for English→Dutch, the **Unsupervised** method is found to outperform even the **Oracle** setting in reducing the amount of

<sup>8</sup>Per-domain breakdown in Figure C.8.

Lang.	MQM Category	MT		No Highlight		Oracle		Unsupervised		Supervised	
		Maj.	Min.	Maj.	Min.	Maj.	Min.	Maj.	Min.	Maj.	Min.
Italian	Accuracy - Addition	0	1	0	0	0	0	0	0	1	1
	Accuracy - Mistranslation	21	22	10	12	4	8	24	17	17	17
	Accuracy - Inconsistency	2	4	1	3	2	2	1	3	0	2
	Accuracy - Omission	2	0	0	0	0	1	4	1	1	2
	Accuracy - Untranslated	1	4	1	2	0	1	1	1	3	2
	Style - Inconsistent Style	0	0	0	0	0	0	0	0	0	0
	Style - Readability	17	25	5	30	0	12	4	34	1	29
	Style - Wrong Register	0	8	0	3	0	3	1	1	3	2
	Linguistic - Grammar	6	15	2	16	0	5	3	12	2	12
	Linguistic - Punctuation	1	13	0	9	0	3	1	6	0	3
	Linguistic - Spelling	5	3	0	4	0	3	3	2	0	1
Dutch	Total	55	95	19	79	6	38	42	77	28	71
	Accuracy - Addition	0	1	0	2	0	3	0	2	0	1
	Accuracy - Mistranslation	25	34	18	25	23	27	12	31	16	29
	Accuracy - Inconsistency	0	0	0	2	0	2	0	2	0	5
	Accuracy - Omission	3	1	1	1	2	1	1	1	4	2
	Accuracy - Untranslated	4	4	1	1	1	4	1	3	0	2
	Style - Inconsistent Style	2	0	0	5	1	7	0	2	0	9
	Style - Readability	1	27	1	20	0	13	2	15	6	41
	Style - Wrong Register	0	2	0	3	0	3	0	1	1	0
	Linguistic - Grammar	3	19	2	14	3	23	2	6	3	12
	Linguistic - Punctuation	0	6	0	3	0	4	0	2	0	3
	Linguistic - Spelling	1	1	1	1	2	1	0	1	0	0
	Total	39	95	24	77	32	88	18	66	30	104

**Table 9.7 |** MQM error counts averaged across  $n = 3$  translators per highlight modality for every translation direction. A description of MQM categories is available in Table 9.1.

errors, while it fares relatively poorly for English→Italian. We also observe a distinct distribution of Accuracy and Style errors, with the formers being more prevalent in biomedical texts, and the latters appearing more frequently in translated social media posts (Figure C.8). We posit that differences in error types across domains might explain the opposite productivity trends observed in Section 9.3.1. While highlighted accuracy errors might lead to time-consuming terminology verification in biomedical texts, style errors might be corrected more quickly and naturally in the social media domain.

**Do Highlights Detect Critical Errors?** We examine whether the critical errors we inserted were detected by different modalities, finding that while most modalities fare decently with more than 62% of critical errors highlighted, **Unsupervised** is the only setting for which all errors are correctly highlighted across both directions (Table 9.8). Then, critical errors are manually verified in all outputs, finding that 16–20% more critical errors are edited in highlighted modalities compared to **No Highlight**. Hence,

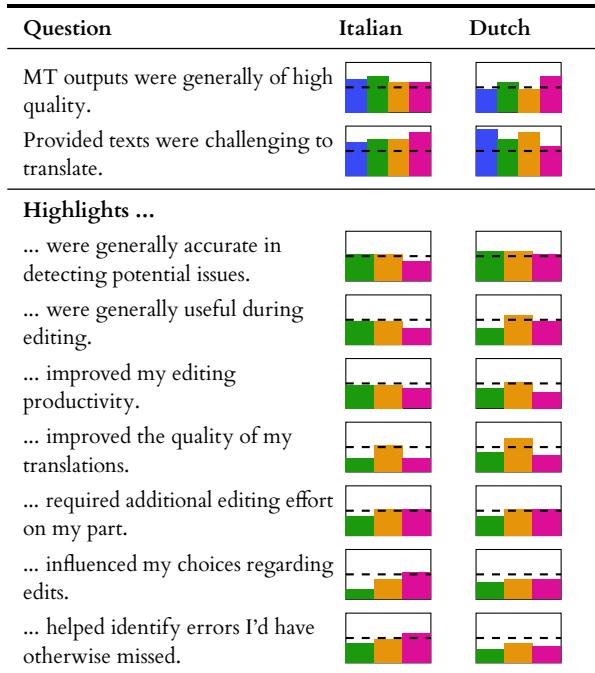
ID	Error Type	Has Highlight			% Post-edited		
		Oracle	Unsup.	Sup.	No High.	Oracle	Unsup.
1-8	Wrong number	NLD	Both	Both	67	83	83
13-6	Remove negation	ITA	Both	Both	50	33	33
16-3	Title literal translation	Both	Both	Both	83	100	100
20-1	Wrong acronym	NLD	Both	ITA	0	33	33
20-7	Wrong acronym (1)	Neither	Both	Neither	0	58	50
20-7	Wrong acronym (2)	NLD	Both	ITA	0	58	50
22-1	Name literal translation	Both	Both	Both	50	50	83
23-4	Addition	NLD	Both	Neither	100	100	83
31-2	Wrong acronym	NLD	Both	Neither	17	33	17
34-7	Numbers swapped	NLD	Both	NLD	17	50	33
37-4	Verb polarity inverted	Both	Both	Both	67	83	67
43-5	Wrong name	Both	Both	Both	50	83	67
48-5	Wrong term	NLD	Both	NLD	67	50	83
<b>Total</b>		65	100	62	44	63	60

**Table 9.8 |** Highlighting and post-editing statistics for manual critical errors (Section 9.2.3). Labels in **Has Highlight** columns indicate whether the error was highlighted in **Both**, only one (**ITA** / **NLD**) only or **Neither** directions. Total scores represent the percentage of detected errors (13 errors, 6 editors per highlight modality).

highlights might lead to narrow but tangible quality improvements that can go undetected in coarse quality assessments, and finer-grained evaluations might be needed to quantify future improvements in word-level QE quality.

### 9.3.4 Usability

In the post-task questionnaire answers (Table 9.9), most translators stated that the MT outputs had an average-to-high quality and that the provided texts were challenging to translate. Highlights were generally found to be reasonably accurate, but they were generally not found helpful to improve either productivity or quality (including **Oracle** ones). Interestingly, despite the convincing gains for critical errors measured in the last section, most translators stated that highlights did not influence their editing and did not help them identify errors that would have otherwise been missed. Specifically, this suggests that translators may not readily perceive potential quality improvements and may consider them secondary to the extra cognitive load elicited by highlighted spans. When asked to comment about highlights, several translators called them “*more of an eye distraction, as they often weren’t actual mistakes*” and “*not quite accurate enough to rely on them as a suggestion*”. Some translators also stated that missed errors led them to



**Table 9.9 |** Post-task questionnaire responses. Bars represent responses ranging from 1–Strongly disagree (no bar) to 5–Strongly agree (full bar), averaged across  $n = 3$  translators per language for No HIGHLIGHT, ORACLE, UNSUPERVISED, and SUPERVISED. Dotted line mark avg. judgments of 3–Neither agree nor disagree.

*“disregarding the highlights to focus on checking each sentence”*. Despite their high quality, only one editor working with Oracle highlights found highlights helpful in *“making the editing process faster and somehow easier”*. Taken together, these comments convincingly suggest a negative perception of the quality and usefulness of highlights, indicating that **improvement in QE accuracy may not be sufficient to enhance QE usefulness** in editors’ eyes.

#### 9.4 Limitations

Our study presents certain limitations that warrant consideration when interpreting its findings and guiding future research.

Firstly, while we included two domains and translation directions to improve the

generalizability of our findings, our results suggest that language and domain play a crucial role in determining the effectiveness of word-level QE for human post-editing. While we observed mild gains from word-level QE on our tested mid-resourced translation directions (English→Italian and English→Dutch), we expect limited, if any, benefit of such approaches in low-resource languages and domains for which MT systems and QE methods are likely to underperform (Sarti et al., 2022; Zouhar et al., 2024). Furthermore, the domains tested in our study (biomedical and social media posts) provided concrete challenges in the form of specialized terminology and idiomatic expressions, respectively, which are known to hinder the quality of MT outputs (Neves et al., 2024; Bawden and Sagot, 2023). While future work should ensure that our findings can be extended to other domains and languages, the limited benefits brought by the tested word-level QE methods in challenging settings suggest a limited usefulness for higher-resource languages and more standard domains, such as news or Wikipedia texts.

Secondly, we acknowledge that several design choices in our evaluation setup, rather than about the QE methods themselves, may have influenced our results. These include, for instance, the specific procedure for discretizing continuous scores from the **Unsupervised** method into error spans, and the method of obtaining **Oracle** highlights via majority voting among post-editors. While we believe these choices are justified within the context of our study, their impact on the outcomes cannot be entirely discounted. Future studies might benefit from a more fine-grained assessment of how such low-level decisions influence the perceived accuracy and usability of word-level QE.

Finally, subjective factors such as the translators' inherent propensity to edit, their prior opinions on the role of MT in post-editing, and their individual editing styles inevitably influenced both quantitative and qualitative assessments in this study. We attempted to mitigate these effects by ensuring a uniform and controlled evaluation setup for all 42 professional translators and by employing averaged judgments for translators working on the same highlight modality. However, we acknowledge that post-editor subjectivity might limit the reproducibility of our findings.

## 9.5 Broader Impact and Ethical Considerations

Our study explicitly centers the experience of professional translators, responding to recent calls for user-centered evaluation of translation technologies. By prioritizing translators' perspectives and productivity, we aim to contribute to methods that complement rather than replace human expertise. Our findings highlight a gap between user perception and measured quality improvements, suggesting that future efforts

should focus primarily on improving the usability of these methods in editing interfaces. In particular, new assistive approaches for post-editing should not only strive to increase productivity but also reduce the cognitive burden associated with post-editing work. This insight is crucial for designing more user-centered quality estimation tools that genuinely support human work. Ultimately, our results suggest that subjective norms across different domains and cultures play an important role in determining the effectiveness of proposed methodologies, underscoring the importance of accounting for human factors when designing such evaluations. All participants in this study were professional translators who provided informed consent. The research protocol ensured anonymity and voluntary participation, with translators recruited and remunerated through professional translation providers. The study's open data release further promotes transparency, enabling other researchers to reproduce and build upon our findings.

## 9.6 Conclusion

Our QE4PE study evaluated the impact of various error-span highlighting modalities, including automatic and human-made ones, on the productivity and quality of human post-editing in a realistic professional setting. Our findings highlight the importance of domain, language and editors' speed in determining the effect of highlights on productivity and quality, underscoring the need for broad evaluations encompassing diverse settings. The limited gains of human-made highlights over automatic QE and their indistinguishable perception from editors' assessment indicate that further gains in the accuracy of these techniques might not be the determining factor in improving their integration into post-editing workflows. In particular, future work might explore other directions to further assess and improve the usability of word-level QE highlights, for example, by studying their impact on non-professional translators and language learners, or by combining them with edit suggestions to justify the presence of error spans.

While our QE4PE study included a preliminary assessment of unsupervised methods for word-level QE, the next chapter expands our investigation by conducting a broad evaluation of unsupervised word-level QE methods for post-editing, including a comparison with state-of-the-art supervised approaches. In light of our results showing the user-dependent effectiveness of quality estimation, we also inspect how the presence of multiple human annotations influences the quality assessment of identified error spans.

# CHAPTER 10

---

## Unsupervised MT Error Detection and Human Disagreement

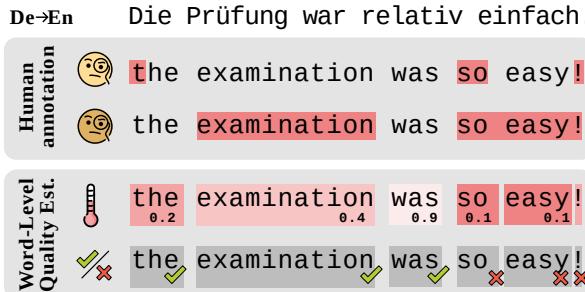
### Chapter Summary

This final experimental chapter presents our comprehensive evaluation of unsupervised word-level quality estimation methods exploiting interpretability and uncertainty quantification methods to identify translation errors in model outputs. In our evaluation spanning 14 metrics across 12 translation directions, we also quantify the impact of human label variation on metric performance, using multiple edit sets from the DivEMT and QE4PE studies of the previous chapters. Our results highlight the untapped potential of unsupervised metrics, the shortcomings of supervised methods when faced with label uncertainty, and the brittleness of single-annotator evaluation practices.

This chapter is adapted from the paper *Unsupervised Word-level Quality Estimation for Machine Translation Through the Lens of Annotators (Dis)agreement* (Sarti et al., 2025a).

*So, you see, translators do not so much deliver a message as they rewrite the original. And herein lies the difficulty—rewriting is still writing, and writing always reflects the authors ideology and biases.*

– Rebecca F. Kuang, *Babel* (2022)



**Figure 10.1** | Example of German→English translation with two sets of human word-level error span annotations and two examples of continuous and binary WQE metrics.

## 10.1 Introduction

Word-level error spans are widely used in machine translation evaluation to obtain robust and fine-grained estimates of translation quality (Lommel et al., 2013; Freitag et al., 2021a,b; Kocmi et al., 2024b). Due to the cost of manual annotation, word-level quality estimation (WQE) was proposed for assisting in annotating error spans over MT outputs (Zouhar et al., 2025). Modern WQE approaches generally rely on costly inference with large language models or ad-hoc training with large amounts of human-annotated texts (Fernandes et al., 2023a; Kocmi and Federmann, 2023b; Guerreiro et al., 2024), making them impractical for less resourced settings (Zouhar et al., 2024).

To improve the efficiency of MT quality assessment, several works explored the use of signals derived from the internals of neural MT systems (Fomicheva et al., 2020, 2021; Leiter et al., 2024), for identifying problems in MT outputs, such as hallucinations (Guerreiro et al., 2023a,b; Dale et al., 2023a,b; Himmi et al., 2024). However, previous works have focused on sentence-level metrics for overall translation quality and do not evaluate performance on multiple label sets due to high annotation costs (Fomicheva et al., 2022b; Zerva et al., 2024).

In this chapter, we conduct a more comprehensive evaluation spanning 10 unsupervised metrics derived from models' inner representations and predictive distributions to identify translation errors at the word level. We test three open-source multilingual MT models and LLMs of varying sizes across 12 translation directions, including typologically diverse languages and challenging textual domains. Importantly, we focus on texts with *multiple* human annotations to measure the impact of individual annotator preferences on metric performance, setting a “human-level” baseline for the WQE task.

We address the following research questions:

- How accurate are unsupervised WQE metrics in detecting MT errors compared to trained metrics and human annotators?
- Are popular supervised WQE metrics well-calibrated?
- Are the relative performances of WQE metrics affected by the variability in human error annotations?

We conclude with recommendations for improving the evaluation and usage of future WQE systems.

## 10.2 Related Work

**Actionable Insights from Interpretability** Advances in interpretability research have elucidated multiple mechanisms underlying decision-making, knowledge representation, and biases in LMs (Ferrando et al., 2024). However, a better understanding of model’s inner workings often did not translate to tangible gains in model design and other practical applications, which remain rarely explored (Mosbach et al., 2024). Some examples in this direction include using targeted machine unlearning methods for safety-critical scenarios (Barez et al., 2025), or the use of attribution for trustworthy context citations in LM generations (Cohen-Wang et al., 2024; Sarti et al., 2024c; Qi, Sarti, et al., 2024). In this study, unsupervised metrics extracted from an MT model during generation are employed to detect errors in models’ generated outputs, following the unsupervised QE paradigm introduced in Section 2.6. This can be seen as a variant of out-of-distribution detection in signal processing research (Hendrycks and Gimpel, 2017).

**Uncertainty Estimation for Language Models** The estimation of uncertainty in language models has garnered increasing attention (Baan et al., 2023), particularly in the context of generation tasks for which the set of plausible responses is large (Giulianelli et al., 2023). Predictive uncertainty is typically decomposed into its *aleatoric* and *epistemic* components, representing respectively the irreducible variability in the modeled phenomena, and the improvable confidence in model predictions (Kiureghian and Ditlevsen, 2009). Popular methods for uncertainty estimation involve the calibration of predictive probabilities to reflect aleatoric uncertainty (Jiang et al., 2020; Ulmer et al., 2022; Zhao et al., 2023; Chen et al., 2023), and conformal sets prediction (Zerva and Martins, 2024; Ravfogel et al., 2023). In this study, we utilize uncertainty signals from the predictive distribution of MT models and their internal processing to efficiently predict the resulting generation quality at a fine-grained, token-level scale.

**Human Label Variation** Human label variation is a type of uncertainty that arises from the inherent variability in human judgments (Plank et al., 2014; Plank, 2022), which can be hard to disentangle from actual annotation mistakes (Snow et al., 2008; Weber-Genzel et al., 2024). The use of multiple references was recently recommended to ensure a sound evaluation of generative LMs, reflecting human-plausible levels of variability (Giulianelli et al., 2023), contrary to standard practices that employ a single set of “gold” labels. In our analysis of QE4PE data, which contains multiple edits, we adopt a perspectivist approach<sup>1</sup> to ensure a robust assessment of WQE metrics by accounting for annotators’ disagreement (Uma et al., 2021).

### 10.3 Models and Datasets

We use datasets containing error annotations or post-edits on the outputs of open-source models to extract unsupervised WQE metrics using real model outputs, thereby avoiding potential confounders. We select the following datasets, summarized in Table 10.1:

**DivEMT** We reuse the DivEMT dataset, introduced in Chapter 8, including out-of-English machine translations towards six typologically diverse target languages (English→Arabic, Italian, Dutch, Turkish, Ukrainian, Vietnamese) produced by Google Translate and mBART-50 1-to-many for a subset of Wiki texts from the FLORES dataset (Goyal et al., 2022), with edits made by professional translators. In this study, we evaluate unsupervised metrics on the mBART-50 1-to-many model, converting the human post-edits into token-level labels to perform a cross-lingual comparison over a fixed set of examples.

**WMT24** The WMT24 dataset is taken from the General Machine Translation Shared Task at WMT 2024 (Kocmi et al., 2024a). It contains evaluation of several machine translation systems across English→{Czech, Hindi, Japanese, Chinese, Russian} (634 segments per language) and Czech→Ukrainian (1954 segments). The human evaluation was conducted using the Error Span Annotation protocol (ESA, Kocmi et al. (2024b)), which involves human annotators highlighting erroneous spans in the translation and marking them as either MINOR or MAJOR errors. This dataset covers the *news*, *social*, and *speech* (with automatic speech recognition) domains. We adopt the official prompting setup from the WMT24 campaign, using the Aya23 model alongside the provided prompt and three in-context translation examples per language to ensure uniformity with previous results.<sup>2</sup> Aya23 is a large language model introduced by Aryabumi et al.

---

<sup>1</sup><https://pdai.info/>

<sup>2</sup><https://github.com/wmt-conference/wmt-collect-translations>

	DivEMT	WMT24	QE4PE
<b>Languages</b>	EN→AR, IT, NL, TR, UK, VI	EN→JA, ZH, HI, CS, RU CS→UK	EN→IT, NL
<b>Errors type</b>	Post-edit	Annotation	Post-edit
<b>Label sets</b>	1	1	6
<b>Domains</b>	Wiki	Multiple	Social, Biomed
<b>MT Model</b>	mBART-50	Aya23	NLLB
<b># Segments</b>	2580	5124	3888

**Table 10.1 |** Summary of tested datasets. Error spans are obtained from explicit error annotations or post-edited spans.

(2024) to improve the multilingual capabilities of the original Aya model (Üstün et al., 2024) on a selected set of 23 languages. The model was included in the WMT24 evaluation by Kocmi et al. (2024a), resulting in the best translation performance among the tested open-source models. The model is a decoder-only transformer model with 40 layers, a model dimension of 8196 and 64 attention heads per layer. Using WMT24 allows us to extend our evaluation to a state-of-the-art LLM, given the popularity of such systems in MT (Kocmi et al., 2023).

**QE4PE** The QE4PE dataset, introduced in Chapter 9, was created to measure the effect of word-level error highlights when included in real-world human post-editing workflows. The QE4PE data provides granular behavioral metrics to evaluate the speed and quality of post-editing of 12 annotators for En→It and En→NL across two challenging textual domains (social posts and biomedical abstracts) and four error span highlighting modalities, including the unsupervised Surprisal MCD<sub>var</sub> method and the supervised XCOMET-XXL we also test in this study. Provided that the presence of error span highlights was found to influence the editing choices of human editors, we limit our evaluation to the six human annotators per language that post-edited sentences without any highlights (3 for the *Oracle Post-edit* task to produce initial human-based highlights, and 3 for the *No Highlight* modality in the main task). This prevents us from biasing our evaluation of WQE metrics in favor of the metrics that influenced editing choices. As for DivEMT, we use the post-edits over translations—in this case, those of the NLLB 3.3B model (NLLB Team et al., 2024)—to produce token-level error spans, enabling an evaluation of WQE metrics across multiple annotation sets.

## 10.4 Evaluated Metrics

The following metrics were evaluated using the Inseq library introduced in Chapter 3.

**Predictive Distribution Metrics** We use the **Surprisal** of the predicted token  $t^*$ , as negative log-probability  $-\log p(t_i^*|t_{<i})$ , and the **Entropy**  $H$  of the output distribution  $P_N$  over vocabulary  $\mathcal{V}$ ,  $-\sum_{i=1}^{|\mathcal{V}|} p(t_i|t_{<i}) \log_2 p(t_i|t_{<i})$ , as simple metrics to quantify pointwise and full prediction uncertainty (Fomicheva et al., 2020). For surprisal, we also compute its expectation (**MCD<sub>avg</sub>**) and variance (**MCD<sub>var</sub>**) with  $n = 10$  steps of Monte Carlo Dropout (MCD, Gal and Ghahramani, 2016) to obtain a robust estimate and a measure of epistemic uncertainty in predictions, respectively. Intuitively, epistemic uncertainty reflects models’ lack of knowledge rather than data ambiguity.<sup>3</sup> We employ the mean of the negative log probabilities as a robust estimate of surprisal:

$$\text{Surprisal MCD}_{\text{avg}} = \hat{y}_{\text{MCD}} = \frac{1}{T} \sum_{t=1}^T -\log p(x|\Theta_t)$$

Moreover, we estimate predictive uncertainty by calculating the variance of predictive probabilities under the same setup:

$$\text{Surprisal MCD}_{\text{var}} = \frac{1}{T} \sum_{t=1}^T (-\log p(x|\Theta_t) - \hat{y}_{\text{MCD}})$$

**Vocabulary Projections** We use the Logit Lens method (LL, nostalgebraist, 2020), introduced in Section 2.1.3, to extract probability distributions  $P_0, \dots, P_{N-1}$  over  $V$  from intermediate activations at every layer  $l_0, \dots, l_{N-1}$  of the decoder. We use the surprisal for the final prediction at every layer (**LL-Surprisal**) to assess the presence of layers with high sensitivity to incorrect predictions. For the NLLB and mBART-50 models, we also apply a final layer normalization before the projection, following the model architecture. For the Aya model, we instead scale logits by 0.0625 (the default `logit_scale` defined in the model configuration). Following the residual stream view of the transformer model (Elhage et al., 2021), the resulting logits offer insight into the model’s predictive confidence at that specific depth of processing. Then, we compute the KL divergence between every layer distribution and the final distribution  $P_N$ , e.g.  $\text{KL}(P_{N-1}\|P_N)$ , to highlight trends in the shift in predictive probability produced

---

<sup>3</sup>MCD is tested only on encoder-decoder models since Aya layers do not include dropout. The  $\text{MCD}_{\text{var}}$  setting corresponds to the **Unsupervised** setting from Chapter 9.

by the application of remaining layers (**LL KL-Div**). Finally, we adapt the approach of Baldock et al. (2021) and use the number of the first layer for which the final prediction corresponds to the top logit as a metric of model confidence,  $l$  s.t.  $\arg \max P_l = t^*$  and  $\arg \max P_i \neq t^* \forall i < l$  (**LL Pred. Depth**).

**Context mixing** We employ simple estimates of context relevance using attention weights produced during the transformer attention operation. More specifically, for every attention head at every layer of the decoder module, we extract a score for every token in the preceding context. We then use the entropy of the distribution of attention weights<sup>4</sup> over previous context as a simple measure of information locality during inference (Ferrando et al., 2022b; Mohebbi et al., 2023). Following Fomicheva et al. (2020), we experiment with using the mean and the maximum entropy across all attention heads of all layers as separate metrics (**Attn. Entropy**<sub>avg/max</sub>). Finally, we evaluate the Between Layer OOD method by Jelenić et al. (2024), employing gradients to estimate layer transformation smoothness for OOD detection (**BLOOD**).

**Supervised baselines** We also test the state-of-the-art supervised WQE model **xcomet** (Guerreiro et al., 2024), introduced in Section 2.6. In this chapter, we focus on their word-level error span prediction capabilities in a quality estimation setup, where the model classifies every input token according to MQM severity levels {OK, MINOR, MAJOR, CRITICAL} with a learned linear layer.<sup>5</sup> Contrary to the continuous metrics from the previous section, binary labels from **xcomet** cannot be easily calibrated to match subjective annotation propensity. Hence, we propose to adapt the **xcomet** metric to use the sum of probability for all error types as a token-level continuous confidence metric,  $s(t^*) = p(\text{minor}) + p(\text{major}) + p(\text{critical})$ , which we dub **xcomet**<sub>conf</sub>.

**Human Editors** For QE4PE, we report the min/mean/max agreement between each annotator’s edited spans and those of the other five editors as a less subjective “human-level” quality measure.

## 10.5 Experiments

### 10.5.1 Setup

**Token-level Evaluation** Error spans used as labels in our evaluation are defined at the character level, while metric scores depend on the tokenization employed by

---

<sup>4</sup>For encoder-decoder model, self-attention and cross-attention weights are concatenated and renormalized.

<sup>5</sup>The default **xcomet** metric was used with the `unbabel-comet` library (v2.2.6).

Source <sub>EN</sub> MT <sub>IT</sub> (NLLB)	So why is it that people jump through extra hoops to install Google Maps? Quindi perché le persone devono fare un salto in più per installare Google Maps?
Annotator <i>t1</i>	Quindi perché le persone devono fare un <b>passaggio</b> in più per installare Google Maps? <small>devono fare un salto in più</small>
Annotator <i>t2</i>	Quindi perché le persone <b>fanno i salti mortali</b> per installare Google Maps? <small>devono fare un salto in più</small>
Annotator <i>t3</i>	Quindi perché le persone <b>effettuano dei passaggi ulteriori e superflui</b> per installare Google Maps? <small>Quindi devono fare salto</small>
Annotator <i>t4</i>	<b>Allora</b> perché le persone <b>fanno</b> un <b>passaggio</b> in più per installare Google Maps? <small>Quindi perché le persone devono fare un salto in più</small>
Annotator <i>t5</i>	<b>E allora mi chiedo: perché gli utenti iPhone si affannano tanto</b> per installare Google Maps? <small>devono fare un salto in più</small>
Annotator <i>t6</i>	Quindi perché le persone <b>fanno di tutto</b> per installare Google Maps? <small>2 1 5 4 6 4</small>
Edit Counts	Quindi perché le persone <b>devono fare</b> un <b>salto</b> in più per installare Google Maps?
xcomet-XL	Quindi perché le persone <b>devono fare</b> un <b>salto in più</b> per installare Google Maps? <small>minor minor</small>
xcomet-XXL	<b>Quindi perché</b> le persone <b>devono fare</b> un <b>salto in più</b> per installare Google Maps? <small>.41 .36 .51 .50 .69 .73 .51 .81 .74 .76 .39 .47 .53 .26</small>
xcomet-XL <sub>CONF</sub>	<b>Quindi</b> perché le persone <b>devono</b> <b>fare</b> un <b>salto</b> in <b>più</b> per <b>install</b> <b>are</b> Google <small>.36 .24</small>
	Maps ? <small>.51 .83 .20 .20 .42 .84 .90 .95 .86 .78 .03 .00 .01 .00</small>
xcomet-XXL <sub>CONF</sub>	<b>Quindi</b> perché le persone <b>devono</b> <b>fare</b> un <b>salto</b> in <b>più</b> per <b>install</b> <b>are</b> Google <small>00 .00</small>
	Maps ? <small>.05 .01 .04 .00 .41 .09 .04 .59 .00 .12 .00 .00 .00 .00</small>
Surprisal MCD <sub>VAR</sub>	Quindi perché le persone <b>devono</b> <b>fare</b> un <b>sal</b> to <b>in</b> <b>più</b> per <b>installare</b> Google <small>.00 .00</small>
	Maps ?

**Table 10.2 |** Annotated example from the En→It portion of the QE4PE dataset. **Top:** Annotator edits with highlighted final text and replaced text on top, with count-based aggregation showing inter-annotator agreement. **Bottom:** Word-level annotations for best-performing metrics discussed in the study.

either the MT model (for unsupervised metrics) or xcomet (for supervised metrics). To facilitate comparison, we label tokens as part of an error span if at least one character contained within them was marked as an error or edited by an annotator. Table 10.2 and Table 10.3 provide examples of various segmentations for the same MT output.

**Constraining generation** Evaluating metrics at the word level can be challenging due to the need for perfect uniformity between model generations and annotated spans. For this reason, we extract unsupervised metrics during generation while force-decoding the annotated outputs from the MT model to ensure perfect adherence with

Source <sub>EN</sub>	So the challenges in this are already showing themselves. I'm likely going to have a VERY difficult time getting a medical clearance due to the FAA's stance on certain medications.													
MT <sub>Cs</sub> (Aya23)	Takže problémy s tím se již projevují. Pravděpodobně budu mít PŘESNĚ obtížný čas dostat lékařské potvrzení kvůli postoji FAA k některým lékům.													
Annotator	Takže problémy s tím se již projevují. Pravděpodobně budu mít PŘESNĚ obtížný čas dostat lékařské potvrzení kvůli postoji FAA k některým lékům.													
xCOMET-XL	Takže problémy s tím se již projevují. Pravděpodobně budu mít PŘESNĚ obtížný čas dostat lékařské potvrzení kvůli postoji FAA k některým lékům.													
xCOMET-XXL	Takže problémy s tím se již projevují. Pravděpodobně budu mít PŘESNĚ obtížný čas dostat lékařské potvrzení kvůli postoji FAA k některým lékům.													
xCOMET-XL CONF	PŘESNĚ obtížný čas dostat lékařské potvrzení kvůli postoji FAA k některým lékům. Takže problémy s tím se již projevují. Pravděpodobně budu mít PŘESNĚ obtížný čas dostat lékařské potvrzení kvůli postoji FAA k některým lékům.	0.23	0.28	0.26	0.28	0.17	0.19	0.31	0.23	0.40	0.48	0.79		
xCOMET-XXL CONF	Takže problémy s tím se již projevují. Pravděpodobně budu mít PŘESNĚ obtížný čas dostat lékařské potvrzení kvůli postoji FAA k některým lékům. Takže problémy s tím se již projevují. Pravděpodobně budu mít PŘESNĚ obtížný čas dostat lékařské potvrzení kvůli postoji FAA k některým lékům.	0.25	0.24	0.26	0.31	0.29	0.23	0.26	0.01	0.00	0.00	0.00	0.00	0.00
Out. Entropy	Takže problémy s tím se již projevují. Pravděpodobně budu mít PŘESNĚ obtížný čas dostat lékařské potvrzení kvůli postoji FAA k některým lékům. Takže problémy s tím se již projevují. Pravděpodobně budu mít PŘESNĚ obtížný čas dostat lékařské potvrzení kvůli postoji FAA k některým lékům.	0.88	1.93	1.88	0.84	1.66	1.13	0.89	0.44	0.22	0.09	2.09		
		3.70	0.09	1.40	1.02	0.64	0.69	0.24	0.80	1.01	0.55	0.18	0.11	

**Table 10.3 |** Annotated example from the En→Cs portion of the WMT24 dataset. **Top:** Annotator edits with highlighted Error Span Annotation of minor and major errors. **Bottom:** Word-level annotations for best-performing metrics discussed in the study.

annotated error spans. In general, such an approach could introduce a problematic confounder in the evaluation, as observed results may be the product of constraining a model towards an unnatural generation, rather than reflecting the underlying phenomena. However, in this study, we carefully ensure that the generation setup matches exactly the one of previous works where the annotated translations were produced, using the same MT model and the same inputs.<sup>6</sup> Hence, the constraining process serves as a simple assurance of conformity in light of potential discrepancies introduced by different decoding strategies, and does not affect the soundness of our method.

<sup>6</sup>Generation parameters such as sampling temperature are not relevant in this setting, provided that they only alter the selection of the following output token, which we do via force-decoding.

Method	DivEMT		WMT24		QE4PE	
	AP	F1*	AP	F1*	AP	F1*
Random	.34	.50	.05	.09	.17	.27
UNSUPERVISED	Surprisal	.43	.53	.08	.13	.23
	Out. Entropy	.46	.51	.10	.16	.23
	Surprisal MCD <sub>AVG</sub>	.43	.53	-	-	.24
	Surprisal MCD <sub>VAR</sub>	.47	.54	-	-	.26
	LL Surprisal <sub>BEST</sub>	.42	.53	.09	.15	.23
	LL KL-Div <sub>BEST</sub>	.43	.51	.07	.12	.20
	LL Pred. Depth	.39	.51	.06	.12	.20
	Att. Entropy <sub>Avg</sub>	.37	.50	.05	.09	.18
	Att. Entropy <sub>Max</sub>	.34	.50	.05	.09	.16
SUPER.	BLOOD <sub>BEST</sub>	.34	.50	-	-	.28
	XCOMET-XL	.42	.45	.09	.19	.23
	XCOMET-XL <sub>CONF</sub>	.54	.55	.15	.23	.32
	XCOMET-XXL	.43	.41	.09	.20	.22
HUM.	XCOMET-XXL <sub>CONF</sub>	.56	.55	.16	.24	.33
	Hum. Editors <sub>MIN</sub>	-	-	-	-	.24
	Hum. Editors <sub>Avg</sub>	-	-	-	-	.28
	Hum. Editors <sub>MAX</sub>	-	-	-	-	.41
					.32	.47

**Table 10.4 |** Average Precision (AP) and Optimal F1 (F1\*) for metrics across tested datasets. Results are averaged across all languages and annotators, with best unsupervised and overall best results highlighted.

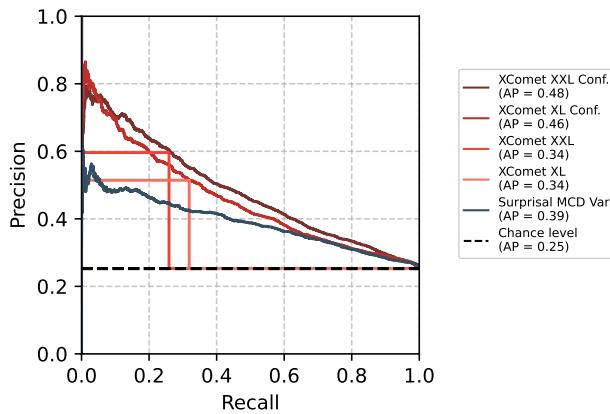
### 10.5.2 Results

**How Accurate are Unsupervised WQE Metrics?** Table 10.4 reports the average metrics performance across all translation directions across tested datasets.<sup>7</sup> We report Average Precision (AP) as a general measure of metric quality across the full score range, and we estimate calibrated metric performance as the best F1 score (F1\*) across all thresholds for binarizing continuous metric scores into pos./neg. labels matching human annotation.<sup>8</sup> Our results show that, despite high variability in error span prevalence across different models, languages and annotators, metric rankings remain generally consistent, suggesting the presence of **robust relations between various signals sourced from models’ inner workings and translation errors**.

Among unsupervised metrics, we find those based on the output distribution to be

<sup>7</sup>Full breakdown available in Table C.16, Table C.17, Table C.18, Table C.19.

<sup>8</sup>Random baseline AP values match the proportion of tokens marked as errors, which can vary greatly.



**Figure 10.2 |** Precision-Recall tradeoff for binary and confidence-weighted xCOMET variants and the Surprisal MCD<sub>var</sub> metric for DivEMT EN→IT.

most effective at identifying error spans, in line with previous segment-level QE results (Fomicheva et al., 2020). Notably, the Surprisal MCD<sub>var</sub> shows strong performances in line with the default xCOMET models. For the multi-label QE4PE dataset, we find that the best supervised metrics score on par with the average human annotator consensus (Hum. Editors<sub>avg</sub>), while unsupervised metrics generally obtain lower performances.

**Confidence Weighting Enables xcomet Calibration** From Table 10.4 results, default xCOMET metrics underperform compared to the best unsupervised techniques, a surprising result given their ad-hoc tuning. On the contrary, simple continuous scores derived from xCOMET ( $xCOMET_{conf}$ ) consistently reach better results across all tested sets. Figure 10.2 shows the precision-recall tradeoff for these metrics on the EN→IT subset of the DivEMT dataset.<sup>9</sup> In their default form, commonly used for evaluation via the `unbabel-comet` library, xCOMET metrics consistently outperform Surprisal MCD<sub>var</sub> in terms of precision (51–60%, compared to 34% optimal precision for MCD<sub>var</sub>), but identify only 32–26% of tokens annotated as errors, resulting in lower AP.

The low recall of these metrics may be problematic in WQE applications, where omitting an error could result in oversights by human post-editors, who may trust the comprehensiveness of WQE predictions. On the contrary, the confidence-weighted  $xCOMET_{conf}$  shows strong performances across the whole recall range, resulting in consistent improvements in both F1\* and AP Table 10.4. Concretely, these results

<sup>9</sup>Results for all datasets in Figure C.11, Figure C.12, Figure C.13, Figure C.14.

confirm that default xCOMET performance does not reflect the full capacity of the metric, and **operating with granular confidence scores can be beneficial when calibration is possible.**

**Metrics Performance for Multiple Annotations** While our evaluation so far employed human error span annotations as binary labels, we set out to assess how more granular labeling schemes impact metrics’ performance. Given  $L$  sets of binary labels (up to 6 per language for QE4PE), we assign a score  $s \in \{1, \dots, L\}$  to every MT token using the number of annotators that marked it as an error, resulting in edit counts reflecting human agreement rate, as shown in Table 10.2.

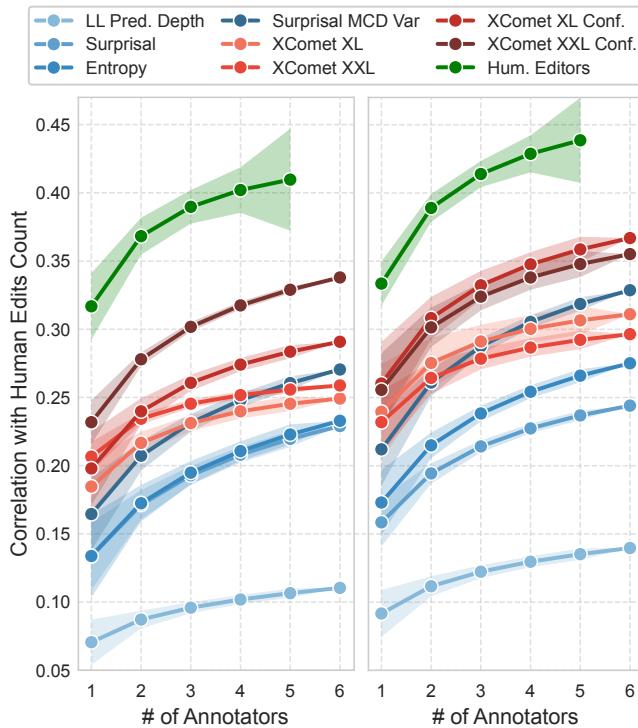
Figure 10.3 presents the correlation of various metrics as the number of annotators available increases, with median values and confidence bounds obtained from edit counts across all combinations of  $L$  label sets.<sup>10</sup> The increasing trend in correlations across all reported metrics indicates that these methods effectively reflect the *aleatoric uncertainty* in error span labels, i.e., the disagreement between various annotators. In particular, the Surprisal MCD<sub>var</sub> metric sees a steeper correlation increase than other well-performing metrics, surpassing default xCOMET supervised approaches for higher correlation bins. This suggests the epistemic uncertainty derived from noisy model predictions might be a promising way to anticipate the aleatoric uncertainty across human annotators for WQE. We observe that 95% confidence intervals for high-scoring metrics largely overlap when a single set of labels is used, indicating that **rankings of metric performance are subject to change depending on the subjective choices of the annotator.** While this poses a problem when attempting a robust evaluation of WQE metrics, we remark that including multiple annotations largely mitigates this issue. As a result, we recommend explicitly accounting for human label variation by including multiple error annotations in future WQE evaluations to ensure generalizable findings.

## 10.6 Limitations

Our findings are accompanied by several limitations. Firstly, our choice of tested datasets was limited by the availability of annotated outputs generated by open-source MT models. While several other datasets matching these criteria exist (Fomicheva et al., 2022b; Yang et al., 2023; Dale et al., 2023b), we restricted our assessment to a sufficient subset to ensure diversity across languages and tested models to support our findings. To facilitate comparison with other datasets, our evaluation for WMT24 treats

---

<sup>10</sup> $x=1$  corresponds to binary labels from previous sections.



**Figure 10.3 |** Spearman correlation between WQE metric scores and human edit counts across multiple annotation sets for QE4PE EN→IT (left) and EN→NL (right).

available error spans as binary labels and does not directly account for error severity in human-annotated spans. Our choice of unsupervised metrics was primarily driven by previous work on uncertainty quantification in MT, and ease of implementation for popular methods in mechanistic interpretability literature (Ferrando et al., 2024). However, our choices in the latter category were limited, as most methods are nowadays developed and tested specifically for decoder-only transformer models. Finally, despite their strong performance, we found that unsupervised methods based on MCD require substantial computational resources, and as such, we were unable to evaluate them on Aya23 35B. While our primary focus was to establish baseline performances across various popular methods, future work should leverage the latest insights from more advanced techniques, such as those requiring the tuning of vocabulary projections (Belrose et al., 2023; Yom Din et al., 2024) or the identification of “confidence neurons” that modulate predictive entropy (Stolfo et al., 2024).

## 10.7 Conclusion

We conducted a comprehensive evaluation of supervised and unsupervised WQE metrics across multiple languages and annotation sets. Our results show that, while unsupervised metrics generally lag behind state-of-the-art supervised systems, some uncertainty quantification methods based on the predictive distribution show promising correlation with human label variation. Moreover, we find that popular supervised WQE metrics generally have low levels of recall and can benefit from confidence weighting when calibration is possible. Finally, individual annotator preferences are key confounders in WQE evaluations and can be mitigated by using multiple annotation sets.

We offer the following practical recommendations for evaluating WQE systems:

- Use agreement between multiple human annotations to control the effect of subjective preferences and rank WQE metrics robustly.
- Employ an in-distribution calibration set of error spans before testing to ensure fair metric comparisons, and favor evaluations accounting for precision-recall tradeoffs to ensure their usability across various confidence levels.
- Previous work showed the effectiveness of visualization reflecting prediction confidence (Vasconcelos et al., 2025), such as highlights for various error severity levels (Sarti et al., 2025b). Consider using continuous WQE metrics in real-world applications such as WQE-augmented post-editing to convey fine-grained confidence variations.

This final assessment concludes our investigation into the potential of model processing signals for enhancing the downstream verification of machine-translated content, converting interpretability methods commonly used for model analysis into practical tools for improving decision-making in real-world human-AI interaction settings.

# CHAPTER 11

---

## Conclusion

*Building a language to communicate with AI isn't going to be easy, but quite frankly, it's the only way to gain control of the way we want to live. Languages shape the way we think. We have an opportunity to shape our own thinking and future machines.*

– Been Kim, *Beyond Interpretability ICLR Keynote (2022)*

Language models have evolved from narrow, task-specific tools to general-purpose architectures that convert knowledge into actionable insights across hundreds of languages. Interpretability research has shed light on how these systems process language, pioneering novel analysis methods to investigate their predictive behaviors and inner mechanisms. Today’s challenge is to translate these insights into practical tools and techniques that help debug models, control their behaviors, and ultimately improve their trustworthiness and usability in the eyes of users. This dissertation has sought to tackle this challenge, developing frameworks that serve the users of language models and machine translation systems at various levels: from everyday users who need factual answers from chatbots, to developers customizing model outputs, to professional editors refining machine translations.

In this final chapter, we begin by revisiting the research questions posed in Chapter 1 and answering them in relation to our findings. We then conclude by charting a path forward, discussing how actionable interpretability research can shape the next generation of transparent, controllable AI systems.

## 11.1 Research Questions Revisited

### ? Research Question 1 (RQ1)

What are the conceptual and technical requirements for interpretability software tools enabling scalable and reproducible analyses into the inner workings of generative language models?

The development and deployment of the Inseq toolkit (Chapter 3) and its subsequent integration with the PECoRe framework (Chapter 4) have provided important insights into this question. From a conceptual standpoint, the main principle to facilitate their widespread adoption is a **progressive disclosure of complexity**, which is necessary to benefit users at all levels of expertise. This human-computer interaction concept proved essential for bridging the gap between two distinct user groups: interpretability researchers with deep technical expertise, and domain experts who understand the practical implications but may lack programming skills. We achieved this balance through three key strategies: first, we unified access to popular models and methods through interfaces compatible with mainstream frameworks. Second, we provided both cutting-edge techniques and extensible baselines with sensible defaults. Third, we created compelling visualizations and post-processing functions that surface key insights without overwhelming users.

On the technical front, supporting model quantization, efficient batching, and distributed inference proved to be challenging yet essential. As language models become increasingly computationally demanding, these optimizations ensure that our tools remain accessible across diverse domains and computational budgets. Our Inseq toolkit successfully innovates across these dimensions, providing simple interfaces for common use cases while maintaining access to advanced features. Its widespread adoption across machine translation, summarization, question answering, and conversational AI validates these design choices and demonstrates their broad applicability.

### ? Research Question 2 (RQ2)

How do language models and machine translation systems exploit contextual information during generation, and how can we quantify this usage in a faithful manner?

Our PECoRE framework (Chapter 4) demonstrated that we can faithfully quantify

context usage in language models and machine translation systems through a two-step process: first, by identifying context-sensitive tokens using contrastive information-theoretic metrics, and then attributing their generation to specific contextual cues through contrastive input attribution. This **data-driven** verification process replaces traditional heuristic-based analyses, enabling model debugging at scale.

Our findings exposed critical weaknesses in context-aware MT systems. In particular, we traced gender agreement failures back to incorrect anaphora resolution and found formatting anomalies triggered by spurious examples in context. When we extended this analysis to retrieval-augmented generation with MIRAGE (Chapter 5), we found that attribution based on model internals could accurately cite relevant retrieved passages. Our proposed procedure avoids the pitfalls of post-hoc rationalizations using surface-level similarity between generated and retrieved contents, grounding instead the citation process in actual context usage for improved trustworthiness.

### ? Research Question 3 (RQ3)

Are interpretability-based steering methods viable approaches for controllable machine translation? How do they compare with prompting-based methods in terms of their performance and their impact on models' internal mechanisms?

The comparative analysis of Chapter 7 established interpretability-based steering as a viable alternative to prompting for controllable machine translation. Our contrastive SAE steering framework matched prompting's personalization accuracy—which already outperformed traditional fine-tuned MT systems in Chapter 6 —while offering distinct advantages in terms of efficiency and transparency. Remarkably, our framework successfully captured individual translators' stylistic signatures using only learned sparse latent representations, succeeding even in the challenging domain of literary translation.

Moreover, our probing analyses revealed that steering and prompting converge on similar mechanistic solutions, resulting in comparable underlying representations. However, steering methods offer crucial advantages: while in-context demonstrations can fail unpredictably based on prompting choices such as example ordering, steering provides direct control through an interpretable concept space with tunable steering intensity.

**? Research Question 4 (RQ4)**

Does MT contribute positively to the productivity of professional translators across different languages? Which factors influence its effectiveness?

The DivEMT study in Chapter 8 provided a nuanced answer to this question. While access to MT generally improved translator productivity, its contribution varied dramatically by language pair. In our results, typological similarity emerged as a significant factor: languages closely related to the source language, English, such as Dutch and Italian, exhibited substantial post-editing productivity gains, whereas distant pairs, like English–Arabic and English–Vietnamese, showed minimal improvement even after controlling for the resourcedness of the training data.

Notably, we found traditional MT quality metrics to correlate poorly with actual productivity benefits across languages. This disconnect challenges the fundamental assumptions that the outcome of better-scoring systems should require less editing, underscoring the need for user-centered assessment that goes beyond technical quality measures.

**? Research Question 5 (RQ5)**

How do word-level error highlights impact the productivity and editing choices of professional translators and the quality of resulting translations?

Our QE4PE study revealed a multifaceted impact of error highlights on the workflow of professional translators. We identified potential error cues that influence both translators' productivity and editing behavior in different ways, with effects that depend heavily on textual domains and translation direction. Interestingly, Italian translators responded to highlights by editing more broadly across entire texts, whereas Dutch translators focused their edits primarily on highlighted spans. These results suggest different approaches to the post-editing task, hinting at the influence of cultural factors at play.

In our error assessment, highlights led to a 15–20% reduction in critical errors compared to standard post-editing, as translators caught mistakes they might otherwise miss. However, overall quality metrics showed no improvement, indicating that coarse-grained quality metrics employed in MT evaluation might fail to capture these targeted benefits. Perhaps most surprisingly, we found no meaningful differences in terms of speed or quality between editors working with human-made highlights, supervised

models, or unsupervised uncertainty metrics. This suggests that the technical accuracy of quality estimation, which is typically the focus of evaluation campaigns, matters less than understanding how to integrate these tools into translators' workflows effectively.

? Research Question 6 (RQ6)

Can unsupervised error span detection methods reliably identify problems in machine-translated outputs? How does human label variation affect their performance, compared to traditional supervised approaches?

Our systematic evaluation in Chapter 10 demonstrated that unsupervised methods employing model internals can match supervised approaches in detecting translation errors across multiple models, datasets, and languages. The variance of token log-probabilities estimated with Monte Carlo Dropout (MCD) proved particularly robust for predicting error spans, outperforming methods based on vocabulary projections, attention weights, and other internal signals.

We found that the limitations of supervised metrics stem from their low recall, with predictions often missing the actual error distribution in test sets. Proper calibration of these metrics' confidence dramatically improved their performance, bringing them close to inter-annotator agreement levels among professional translators. Crucially, we found that metric rankings can shift substantially when few annotations are present, depending on individual annotators' subjective judgments. This underscores the necessity of multiple annotation sets and careful calibration for fair quality estimation assessment.

## 11.2 Outlook and Future Directions

The themes and findings of this dissertation open several promising avenues toward the role of actionable interpretability insights for trustworthy NLP systems.

A core premise of this thesis—that downstream applications serve as invaluable testbeds for interpretability methods—resonates with current debates within the interpretability research community. As Marks (2025) argues, if interpretability methods enable use cases unattainable by other approaches, they provide evidence of genuine, significant insights. Our work validates this perspective: interpretability-based methods excel at answer attribution, controlled generation and error detection, providing more faithful and auditable results than the supervised models typically employed for those tasks. Our PECoRE framework, for example, can debug issues in context usage that would be hard to detect through simple behavioral evaluations.

The final experimental chapters of this thesis take this paradigm a step further, evaluating interpretability techniques not only by their accuracy on realistic tasks but also by their downstream impact on user decision-making, productivity, and satisfaction. While the focus of the NLP interpretability community in recent years has gravitated towards the low-level technical depths of *mechanistic interpretability* (Saphra and Wiegreffe, 2024), the emerging field of human-centered explainable AI (HCXAI)—which has for now mainly engaged the human-computer interaction community<sup>1</sup>—is taking the lead in developing sociotechnical frameworks for model explanations centered around users’ needs and experiences. The intersection between these areas remains frustratingly small: few mechanistic studies conduct downstream human evaluations, and most human-centered work fails to integrate the best state-of-the-art interpretability methods due to a lack of experience or resources. Work aimed at bridging this gap will remain essential to ensure that interpretability advances remain both technically sound and practically relevant.

Despite its success, modern interpretability research faces a serious threat: the *growing inaccessibility* of frontier systems, which play a key role as prime “subjects” of interpretability studies. A recent survey of 184 recent interpretability works reveals a widening disparity between the capabilities of state-of-the-art systems and those of systems generally evaluated in interpretability studies (Fiotto-Kaufman et al., 2025). This gap, driven by engineering barriers and proprietary API restrictions, threatens the validity of insights derived from simpler, less capable models. Addressing this issue will require robust shared infrastructure for interpretability research, simplifying access to state-of-the-art systems and fostering a more inclusive research environment. Our proposed Inseq library was developed with this in mind, supporting methods such as quantized, batched and distributed inference to reduce the computational load of interpretability analyses. More recently, the NNSight library (Fiotto-Kaufman et al., 2025) represents the most significant step in this direction, providing researchers with fine-grained access to model internals through remote execution, abstracting away the complexity and costs associated with local hardware setups. Beyond tools, the computational cost of current interpretability methods is a barrier to their widespread adoption, particularly in production environments where faster predictions might be favored over more precise or trustworthy results. Future technical research should prioritize the development of more efficient techniques, exploring approximation methods, caching strategies, or ad-hoc kernels, while preserving faithfulness to the model’s inner workings. The CAT method from Chapter 3, which approximates patching with

---

<sup>1</sup>The main workshop in this area is organized by the ACM SIGCHI interest group.

contrastive gradient attribution, exemplifies the many possibilities in this direction. Perhaps most importantly, interpretability research can pave the way for more effective human-AI collaboration. Our translator studies show that model insights have the potential to transform professional workflows, but also that presentation matters as much, if not more, than accuracy. Despite that, the presentation of interpretability insights is often overlooked by current work. The translation domain presents unique challenges in this area, with human professionals operating in similar settings but across entirely different languages and cultural contexts, requiring tailored approaches. User-centric interfaces that let domain experts explore model behaviors quickly and intuitively will be essential for addressing these challenges.

As language model adoption accelerates, the demand for transparency and usability tools will only intensify, and interpretability researchers are in a crucial position to address these requests. The methods, insights, and perspectives presented in this dissertation demonstrate the potential of interpretability in machine translation, while highlighting the critical importance of continued research at the intersection of interpretability, multilingual NLP, and human-computer interaction. By making these systems more transparent, controllable, and aligned with human needs, we move toward a future where language technologies do not operate as opaque oracles, but rather as trusted partners helping us tackle the complex challenges ahead.



# Bibliography

- Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Y. Zou. 2019. Gradio: Hassle-free sharing and testing of ML models in the wild. *ArXiv*, abs/1906.02569.
- Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.
- Reduan Achitbat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. 2024. Attnlrp: attention-aware layer-wise relevance propagation for transformers. In *Proceedings of the 41st International Conference on Machine Learning*, ICML’24. JMLR.org.
- Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, pages 9505–9515, Montréal, Canada. Curran Associates, Inc.
- Julius Adebayo, Michael Muelly, Harold Abelson, and Been Kim. 2022. Post hoc explanations may be ineffective for detecting unknown spurious correlation. In *International Conference on Learning Representations*.
- Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. 2020. Debugging tests for model explanations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS’20, Red Hook, NY, USA. Curran Associates Inc.
- Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models. *Arxiv*.
- Sweta Agrawal, António Farinhas, Ricardo Rei, and Andre Martins. 2024. Can automatic metrics assess high-quality translations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14491–14502, Miami, Florida, USA. Association for Computational Linguistics.
- Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

- Roe Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.
- Arafat Ahsan, Vandana Mujadia, and Dipti Misra Sharma. 2021. Assessing post-editing effort in the English-Hindi direction. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 44–53, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).
- J Alammar. 2021. Ecco: An open source library for the explainability of transformer language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 249–257, Online. Association for Computational Linguistics.
- Simone Alghisi, Massimo Rizzoli, Gabriel Roccabruna, Seyed Mahed Mousavi, and Giuseppe Riccardi. 2024. Should we fine-tune or RAG? evaluating different techniques to adapt LLMs for dialogue. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 180–197, Tokyo, Japan. Association for Computational Linguistics.
- Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.
- Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2023. ACES: Translation accuracy challenge sets at WMT 2023. In *Proceedings of the Eighth Conference on Machine Translation*, pages 695–712, Singapore. Association for Computational Linguistics.
- Chantal Amrhein and Rico Sennrich. 2021. How suitable are subword segmentation strategies for translating non-concatenative morphology? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 689–705, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Chantal Amrhein and Rico Sennrich. 2022. Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1125–1141, Online only. Association for Computational Linguistics.

- Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, and 1 others. 2023. Palm 2 technical report. *Arxiv*.
- Dana Arad, Yonatan Belinkov, Hanjie Chen, Najoung Kim, Hosein Mohebbi, Aaron Mueller, Gabriele Sarti, and Martin Tutek. 2025. Findings of the BlackboxNLP 2025 shared task: Localizing circuits and causal variables in language models. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 543–552, Suzhou, China. Association for Computational Linguistics.
- Andy Ardit, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, pages 136037–136083, Red Hook, NY, USA. Curran Associates, Inc.
- Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2018. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495.
- Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, and 2 others. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.
- Akari Asai, Xinyan Yu, Jungo Kasai, and Hanna Hajishirzi. 2021. One question answering model for many languages with cross-lingual dense passage retrieval. *Advances in Neural Information Processing Systems*, 34:7547–7560.
- Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–294, Toronto, Canada. Association for Computational Linguistics.
- Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.
- Giuseppe Attanasio, Eliana Pastor, Chiara Di Bonaventura, and Debora Nozza. 2023. ferret: a framework for benchmarking explainers on transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 256–266, Dubrovnik, Croatia. Association for Computational Linguistics.

- Wilker Aziz, Sheila Castilho, and Lucia Specia. 2012. PET: a tool for post-editing and assessing machine translation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3982–3987, Istanbul, Turkey. European Language Resources Association (ELRA).
- Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *Arxiv Preprint*.
- Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty in natural language generation: From theory to applications. *Preprint*, arXiv:2307.15703.
- Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Kluschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46.
- David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. 2010. How to explain individual classification decisions. *J. Mach. Learn. Res.*, 11:1803–1831.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, USA.
- Robert Baldock, Hartmut Maennel, and Behnam Neyshabur. 2021. Deep learning through the lens of example difficulty. In *Advances in Neural Information Processing Systems*, volume 34, pages 10876–10889. Curran Associates, Inc.
- Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Fazl Barez, Tingchen Fu, Ameya Prabhu, Stephen Casper, Amartya Sanyal, Adel Bibi, Aidan O’Gara, Robert Kirk, Ben Bucknall, Tim Fist, Luke Ong, Philip Torr, Kwok-Yan Lam, Robert Trager, David Krueger, Sören Mindermann, José Hernandez-Orallo, Mor Geva, and Yarin Gal. 2025. Open problems in machine unlearning for ai safety. *Preprint*, arXiv:2501.04952.
- Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors. 2021. *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics, Online.

- Jasmijn Bastings, Sebastian Ebert, Polina Zablotskaia, Anders Sandholm, and Katja Filippova. 2022. “will you find these shortcuts?” a protocol for evaluating the faithfulness of input salience methods for text classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 976–991, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, Online. Association for Computational Linguistics.
- Rachel Bawden and Benoît Sagot. 2023. RoCS-MT: Robustness challenge set for machine translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 198–216, Singapore. Association for Computational Linguistics.
- Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.
- Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.
- Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.
- Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.
- Yonatan Belinkov, Aaron Mueller, Najoung Kim, Hosein Mohebbi, Hanjie Chen, Dana Arad, and Gabriele Sarti, editors. 2025. *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Suzhou, China.
- Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *ArXiv*, abs/2303.08112.
- Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas. Association for Computational Linguistics.

- Nathaniel Berger, Stefan Riezler, Miriam Exel, and Matthias Huck. 2024. Post-edits are preferences too. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1289–1300, Miami, Florida, USA. Association for Computational Linguistics.
- Federico Bianchi, Giuseppe Attanasio, Raphael Pisoni, Silvia Terragni, Gabriele Sarti, and Dario Balestri. 2023. Contrastive language–image pre-training for the Italian language. In *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023)*, pages 78–85, Venice, Italy. CEUR Workshop Proceedings.
- BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Luccioni, François Yvon, and 1 others. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *Arxiv*.
- Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. 2024. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences*, 121(2):e2304406120.
- Alexandra Birch, Miles Osborne, and Philipp Koehn. 2008. Predicting success in machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 745–754, Honolulu, Hawaii. Association for Computational Linguistics.
- Arianna Bisazza, Ahmet Üstün, and Stephan Sportel. 2021. On the difficulty of translating free-order case-marking languages. *Transactions of the Association for Computational Linguistics*, 9:1233–1248.
- Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin. Association for Computational Linguistics.
- Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanodia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. Findings of the WMT 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore. Association for Computational Linguistics.
- John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 315–321, Geneva, Switzerland. COLING.
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in NLP. In *Proceedings of the 58th*

- Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.
- Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roei Aharoni, Daniel Andor, Livio Baldini Soares, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *ArXiv*.
- Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.
- Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.
- Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, and 9 others. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.
- Alexander Borzunov, Dmitry Baranchuk, Tim Dettmers, Maksim Riabinin, Younes Belkada, Artem Chumachenko, Pavel Samygin, and Colin Raffel. 2023. Petals: Collaborative inference and fine-tuning of large models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 558–568, Toronto, Canada. Association for Computational Linguistics.
- Lynne Bowker. 2002. *Computer-Aided Translation Technology: A Practical Introduction*. University of Ottawa Press.
- Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021. Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3199–3216, Online. Association for Computational Linguistics.

- Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024. Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1301–1317, Miami, Florida, USA. Association for Computational Linguistics.
- Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, and 6 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS’20, Red Hook, NY, USA. Curran Associates Inc.
- Emanuele Bugliarello, Sabrina J. Mielke, Antonios Anastasopoulos, Ryan Cotterell, and Naoki Okazaki. 2020. It’s easier to translate out of English than into it: Measuring neural translation difficulty by cross-mutual information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1640–1649, Online. Association for Computational Linguistics.
- Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.
- Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.
- Sara Candussio, Gaia Saveri, Gabriele Sarti, and Luca Bortolussi. 2025. Bridging logic and learning: Decoding temporal logic embeddings via transformers. In *Machine Learning and Knowledge Discovery in Databases. Research Track*, ECML-PKDD. Springer Nature Switzerland.
- Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémie Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, and 2 others. 2024. Black-box access is insufficient for rigorous ai audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’24, page 2254–2272, New York, NY, USA. Association for Computing Machinery.

- Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, 108(1):109–120.
- Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14, Tokyo, Japan. International Workshop on Spoken Language Translation.
- Sviatoslav Chalnev, Matthew Siu, and Arthur Conmy. 2024. Improving steering vectors by targeting sparse autoencoder features. *Arxiv*.
- Yangyi Chen, Lifan Yuan, Ganqu Cui, Zhiyuan Liu, and Heng Ji. 2023. A close look into the calibration of pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1343–1367, Toronto, Canada. Association for Computational Linguistics.
- Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy. Association for Computational Linguistics.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, and 48 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.
- George Chrysostomou and Nikolaos Aletras. 2022. An empirical study on explanations in out-of-domain settings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6920–6938, Dublin, Ireland. Association for Computational Linguistics.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.
- Kenneth W. Church and Eduard H. Hovy. 1993. Good applications for crummy machine translation. *Machine Translation*, 8(4):239–258.

- Cristiano Ciaccio, Gabriele Sarti, Alessio Miaschi, and Felice Dell’Orletta. 2025. Crossword space: Latent manifold learning for italian crosswords and beyond. In *Proceedings of the 11th Italian Conference on Computational Linguistics (CLiC-it 2023)*, Cagliari, Italy. CEUR Workshop Proceedings.
- Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT’s attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.
- Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Mądry. 2024. Contextcite: Attributing model generation to context. In *Advances in Neural Information Processing Systems*, volume 37, pages 95764–95807. Curran Associates, Inc.
- Çagrı Çöltekin and Taraka Rama. 2023. What do complexity measures measure? correlating and validating corpus-based measures of morphological complexity. *Linguistics Vanguard*, 9(s1):27–43.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.
- Sven Coppers, Jan Van den Bergh, Kris Luyten, Karin Coninx, Iulianna Van der Lek-Ciudin, Tom Vanallemeersch, and Vincent Vandeghinste. 2018. Intellingo: An intelligible translation environment. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13.
- Ryan Cotterell, Sabrina J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 536–541, New Orleans, Louisiana. Association for Computational Linguistics.
- Ian Covert, Scott Lundberg, and Su-In Lee. 2021. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90.

- Jonathan Crabbé and Mihaela van der Schaar. 2023. Evaluating the robustness of interpretability methods through explanation invariance and equivariance. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5420–5443, Albuquerque, New Mexico. Association for Computational Linguistics.
- Anna Currey, Maria Nadejde, Raghavendra Reddy Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. 2022. MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4287–4299, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Joke Daems, Sonia Vandepitte, Robert Hartsuiker, and Lieve Macken. 2017a. Translation methods and experience: A comparative analysis of human translation and post-editing with students and professional translators. *Meta : journal des traducteurs / Meta: Translators' Journal*, 62(2):245–270.
- Joke Daems, Sonia Vandepitte, Robert J. Hartsuiker, and Lieve Macken. 2017b. Identifying the machine translation error types with the greatest impact on post-editing effort. *Frontiers in Psychology*, 8.
- Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.
- David Dale, Elena Voita, Loic Barrault, and Marta R. Costa-jussà. 2023a. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity Even better. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–50, Toronto, Canada. Association for Computational Linguistics.
- David Dale, Elena Voita, Janice Lam, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Loic Barrault, and Marta Costa-jussà. 2023b. HalOmi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 638–653, Singapore. Association for Computational Linguistics.
- Xuan-Quy Dao and Ngoc-Bich Le. 2023. Chatgpt is good but bing chat is better for vietnamese students. *Arxiv*.

- Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Gpt3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, pages 30318–30332. Curran Associates, Inc.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.
- David L. Donoho and Michael Elad. 2003. Optimally sparse representation in general (nonorthogonal) dictionaries via  $\ell_1$  minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202.
- Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *Preprint*, arXiv:1702.08608.
- Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.
- Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.
- Lukas Edman, Gabriele Sarti, Antonio Toral, Gertjan van Noord, and Arianna Bisazza. 2024. Are character-level translations worth the wait? comparing ByT5 and mT5 for machine translation. *Transactions of the Association for Computational Linguistics*, 12:392–410.
- Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI ’21, New York, NY, USA. Association for Computing Machinery.

- Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I-Hsiang Lee, Michael Muller, and Mark O Riedl. 2024. The who in xai: How ai background shapes perceptions of ai explanations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.
- Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Transformer Circuits Thread*.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. [Https://transformer-circuits.pub/2021/framework/index.html](https://transformer-circuits.pub/2021/framework/index.html).
- Joseph Enguehard. 2023. Sequential integrated gradients: a simple but effective method for explaining language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7555–7565, Toronto, Canada. Association for Computational Linguistics.
- Sugyeong Eo, Chanjun Park, Hyeyonseok Moon, Jaehyung Seo, and Heuiseok Lim. 2022. Word-level quality estimation for korean-english neural machine translation. *IEEE Access*, 10:44964–44973.
- Johannes Eschbach-Dymanus, Frank Essberger, Bianka Buschbeck, and Miriam Exel. 2024. Exploring the effectiveness of LLM domain adaptation for business IT machine translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 610–622, Sheffield, UK. European Association for Machine Translation (EAMT).
- Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9367–9385, Bangkok, Thailand. Association for Computational Linguistics.
- Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim

- Panov, Timothy Baldwin, and Artem Shelmanov. 2023. LM-polygraph: Uncertainty estimation for language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 446–461, Singapore. Association for Computational Linguistics.
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Çelebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.
- Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.
- Anna Farkas and Renáta Németh. 2022. How to measure gender bias in machine translation: Real-world oriented machine translators, multiple reference points. *Social Sciences & Humanities Open*, 5(1):100239.
- Marcello Federico, Nicola Bertoldi, Marco Trombetti, and Alessandro Cattelan. 2014. Mate-Cat: an open source CAT tool for MT post-editing. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: Tutorials*, Vancouver, Canada. Association for Machine Translation in the Americas.
- Thomas Fel. 2024. *Sparks of Explainability: Recent Advancements in Explaining Large Vision Models*. Ph.D. thesis, University of Toulouse.
- Nils Feldhus, Robert Schwarzenberg, and Sebastian Möller. 2021. Thermostat: A large collection of NLP model explanations and analysis tools. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 87–95, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.
- Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium. Association for Computational Linguistics.
- Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023a. The devil

- is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for Computational Linguistics.
- Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.
- Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023b. When does translation require context? a data-driven, multilingual exploration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 606–626, Toronto, Canada. Association for Computational Linguistics.
- Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing context usage in context-aware machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6467–6478, Online. Association for Computational Linguistics.
- Javier Ferrando, Gerard I. Gállego, Belen Alastruey, Carlos Escolano, and Marta R. Costa-jussà. 2022a. Towards opening the black box of neural machine translation: Source and target interpretations of the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8756–8769, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Javier Ferrando, Gerard I. Gállego, and Marta R. Costa-jussà. 2022b. Measuring the mixing of contextual information in the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8698–8714, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Javier Ferrando, Gerard I. Gállego, Ioannis Tsiamas, and Marta R. Costa-jussà. 2023. Explaining how transformers use context to build predictions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5513, Toronto, Canada. Association for Computational Linguistics.
- Javier Ferrando, Oscar Balcells Obeso, Senthooran Rajamanoharan, and Neel Nanda. 2025. Do i know this entity? knowledge awareness and hallucinations in language models. In *The Thirteenth International Conference on Learning Representations*.
- Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. 2024. A primer on the inner workings of transformer-based language models. *Arxiv Preprint*.

- Jaden Fried Fiotto-Kaufman, Alexander Russell Loftus, Eric Todd, Jannik Brinkmann, Koyena Pal, Dmitrii Troitskii, Michael Ripa, Adam Belfki, Can Rager, Caden Juang, Aaron Mueller, Samuel Marks, Arnab Sen Sharma, Francesca Lucchetti, Nikhil Prakash, Carla E. Brodley, Arjun Guha, Jonathan Bell, Byron C Wallace, and David Bau. 2025. NNsight and NDIF: Democratizing access to open-weight foundation model internals. In *The Thirteenth International Conference on Learning Representations*, ICLR.
- Lucie Flekova, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar, and Daniel Preoțiuc-Pietro. 2016. Analyzing biases in human perception of user age and gender from text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 843–854, Berlin, Germany. Association for Computational Linguistics.
- Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021. The Eval4NLP shared task on explainable quality estimation: Overview and results. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 165–178, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Marina Fomicheva and Lucia Specia. 2019. Taking MT evaluation metrics to extremes: Beyond correlation with human judgments. *Computational Linguistics*, 45(3):515–558.
- Marina Fomicheva, Lucia Specia, and Nikolaos Aletras. 2022a. Translation error detection as rationale extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4148–4159, Dublin, Ireland. Association for Computational Linguistics.
- Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2022b. MLQE-PE: A multilingual quality estimation and post-editing dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4963–4974, Marseille, France. European Language Resources Association.
- Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.
- Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.
- Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

- Markus Freitag, Nitika Mathur, Chi-ku Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.
- Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-ku Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-ku Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.
- Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, NY, USA. Proceedings of Machine Learning Research (PLMR).
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2021. The pile: An 800gb dataset of diverse text for language modeling. *Arxiv*.
- Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023a. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023b. Retrieval-augmented generation for large language models: A survey. *ArXiv*.
- Ignacio Garcia. 2009. Beyond translation memory: Computers and the professional translator. *The Journal of Specialised Translation*.
- Xavier Garcia, Noah Constant, Mandy Guo, and Orhan Firat. 2021. Towards universality in multilingual text rewriting. *Arxiv*.

- Xavier Garcia and Orhan Firat. 2022. Using natural language prompts for machine translation. *Arxiv*.
- Xiao Ge, Chunchen Xu, Daigo Misaki, Hazel Rose Markus, and Jeanne L Tsai. 2024. How culture shapes what people want from ai. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.
- Viveta Gene. 2021. The post-editing workflow: Training challenges for LSPs, post-editors and academia. In *Proceedings of the Translation and Interpreting Technology Online Conference*, pages 187–198, Held Online. INCOMA Ltd.
- Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 316–327, Brussels, Belgium. Association for Computational Linguistics.
- Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Mario Julianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What comes next? evaluating uncertainty in neural text generators against human production variability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14349–14371, Singapore. Association for Computational Linguistics.
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.
- Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 29–33, Online. Association for Computational Linguistics.
- Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

- Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.
- Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.
- Spence Green, Jeffrey Heer, and Christopher D. Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’13, pages 439–448, New York, NY, USA. Association for Computing Machinery.
- Ana Guerberof. 2009. Productivity and quality in MT post-editing. In *Beyond Translation Memories: New Tools for Translators Workshop*, Ottawa, Canada.
- Ana Guerberof-Arenas and Joss Moorkens. 2023. *Ethics and Machine Translation: The End User Perspective*, pages 113–133. Springer International Publishing, Cham.
- Ana Guerberof-Arenas and Antonio Toral. 2022. Creativity in translation: Machine translation as a constraint for literary texts. *Translation Spaces*, 11(2):184–212.
- Nuno M. Guerreiro, Pierre Colombo, Pablo Piantanida, and André Martins. 2023a. Optimal transport for unsupervised hallucination detection in neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13766–13784, Toronto, Canada. Association for Computational Linguistics.
- Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.
- Nuno M. Guerreiro, Elena Voita, and André Martins. 2023b. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics.
- Abhijeet Gupta, Gemma Boleda, Marco Baroni, and Sebastian Padó. 2015. Distributional vectors encode referential attributes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 12–21, Lisbon, Portugal. Association for Computational Linguistics.

- Christian Hadiwinoto. 2017. Book review: Syntax-based statistical machine translation by philip Williams, rico Sennrich, matt post and philipp Koehn. *Computational Linguistics*, 43(4):893–896.
- Zellig S. Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Los Alamitos, CA, USA. IEEE Computer Society.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of the 11th International Conference on Learning Representations*, ICLR.
- Roei Hendel, Mor Geva, and Amir Globerson. 2023. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9318–9333, Singapore. Association for Computational Linguistics.
- Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR 2017)*.
- Dan Hendrycks and Laura Hiscott. 2025. The misguided quest for mechanistic ai interpretability. AI Frontiers. Accessed August 4, 2025.
- Nico Herbig, Tim Düwel, Santanu Pal, Kalliopi Meladaki, Mahsa Monshizadeh, Antonio Krüger, and Josef van Genabith. 2020. MMPE: A Multi-Modal Interface for Post-Editing Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1691–1702, Online. Association for Computational Linguistics.
- Anas Himmi, Guillaume Staerman, Marine Picot, Pierre Colombo, and Nuno M Guerreiro. 2024. Enhanced hallucination detection in neural machine translation through simple detector aggregation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18573–18583, Miami, Florida, USA. Association for Computational Linguistics.
- Sepp Hochreiter. 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 6(2):107–116.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn’t always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

- Or Honovich, Roei Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.
- Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.
- Jing Huang, Atticus Geiger, Karel D’Oosterlinck, Zhengxuan Wu, and Christopher Potts. 2023. Rigorously assessing natural language explanations of neurons. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 317–331, Singapore. Association for Computational Linguistics.
- Lianzhe Huang, Shuming Ma, Dongdong Zhang, Furu Wei, and Houfeng Wang. 2022. Zero-shot cross-lingual transfer of prompt-based tuning with a unified multilingual prompt. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11488–11497, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*.
- William J. Hutchins. 2001. Machine translation over fifty years. *Histoire Épistémologie Langage*, 23:7–31.
- Khondoker Ittehadul Islam and Gabriele Sarti. 2025. Reveal-bangla: A dataset for cross-lingual multi-step reasoning evaluation. *Arxiv Preprint*.
- Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.
- Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.
- Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational*

- Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.
- Stanisław Jastrzebski, Devansh Arpit, Nicolas Ballas, Vikas Verma, Tong Che, and Yoshua Bengio. 2018. Residual connections encourage iterative inference. In *International Conference on Learning Representations*.
- Fran Jelenić, Josip Jukić, Martin Tutek, Mate Puljiz, and Jan Snajder. 2024. Out-of-distribution detection by leveraging between-layer transformation smoothness. In *The Twelfth International Conference on Learning Representations*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.
- Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.
- Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.
- Linghao Jin, Jacqueline He, Jonathan May, and Xuezhe Ma. 2023. Challenges in context-aware neural machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15246–15263, Singapore. Association for Computational Linguistics.
- Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics.
- Jared Kaplan, Sam McCandlish, T. J. Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *ArXiv*.
- Sariya Karimova, Patrick Simianer, and Stefan Riezler. 2018. A user-study on online adaptation of neural machine translation to human post-edits. *Machine Translation*, 32(4):309–324.
- Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics.

- Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy. Association for Computational Linguistics.
- Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019a. Pivot-based transfer learning for neural machine translation between non-English languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 866–876, Hong Kong, China. Association for Computational Linguistics.
- Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019b. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DisCoMT 2019)*, pages 24–34, Hong Kong, China. Association for Computational Linguistics.
- Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112. Risk Acceptance and Risk Communication.
- Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.
- Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2021. Incorporating Residual and Normalization Layers into Analysis of Masked Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4547–4568, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- 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, and 3 others. 2024a. 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.
- Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and 3 others. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

- Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.
- Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.
- Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024b. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.
- Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.
- Arne Köhn. 2015. What's in an embedding? analyzing word embeddings through multilingual evaluation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2067–2073, Lisbon, Portugal. Association for Computational Linguistics.
- Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for pytorch. *ArXiv*.
- Maarit Koponen, Wilker Aziz, Luciana Ramos, and Lucia Specia. 2012. Post-editing time as a measure of cognitive effort. In *Workshop on Post-Editing Technology and Practice*.
- Maarit Koponen, Umut Sulubacak, Kaisa Vitikainen, and Jörg Tiedemann. 2020. MT for subtitling: User evaluation of post-editing productivity. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 115–124, Lisboa, Portugal. European Association for Machine Translation.
- Hans P. Krings. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. Kent State University Press.
- Kalpesh Krishna, Deepak Nathani, Xavier Garcia, Bidisha Samanta, and Partha Talukdar. 2022. Few-shot controllable style transfer for low-resource multilingual settings. In

- Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7439–7468, Dublin, Ireland. Association for Computational Linguistics.
- Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957, Online. Association for Computational Linguistics.
- Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. 2024. The disagreement problem in explainable machine learning: A practitioner’s perspective. *Transactions on Machine Learning Research*.
- Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.
- Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.
- Isabel Lacruz, Michael Denkowski, and Alon Lavie. 2014. Cognitive demand and cognitive effort in post-editing. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, pages 73–84, Vancouver, Canada. Association for Machine Translation in the Americas.
- Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. An evaluation of the human-interpretability of explanation. *ArXiv*, abs/1902.00006.
- Huiyuan Lai, Jiali Mao, Antonio Toral, and Malvina Nissim. 2022. Human judgement as a compass to navigate automatic metrics for formality transfer. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 102–115, Dublin, Ireland. Association for Computational Linguistics.
- Surafel Melaku Lakew, Mattia Di Gangi, and Marcello Federico. 2019. Controlling the output length of neural machine translation. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.
- Anna Langedijk, Hosein Mohebbi, Gabriele Sarti, Willem Zuidema, and Jaap Jumelet. 2024. DecoderLens: Layerwise interpretation of encoder-decoder transformers. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4764–4780, Mexico City, Mexico. Association for Computational Linguistics.

- Samuel Läubli, Chantal Amrhein, Patrick Düggelin, Beatriz Gonzalez, Alena Zwahlen, and Martin Volk. 2019. Post-editing productivity with neural machine translation: An empirical assessment of speed and quality in the banking and finance domain. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 267–272, Dublin, Ireland. European Association for Machine Translation.
- Samuel Läubli, Mark Fishel, Gary Massey, Maureen Ehrensberger-Dow, and Martin Volk. 2013. Assessing post-editing efficiency in a realistic translation environment. In *Proceedings of the 2nd Workshop on Post-editing Technology and Practice*, Nice, France.
- 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.
- Jihyeon Lee, Taehee Kim, Yunwon Tae, Cheonbok Park, and Jaegul Choo. 2023a. PePe: Personalized post-editing model utilizing user-generated post-edits. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 239–253, Dubrovnik, Croatia. Association for Computational Linguistics.
- Seungjun Lee, Jungseob Lee, Hyeonseok Moon, Chanjun Park, Jaehyung Seo, Sugyeong Eo, Seonmin Koo, and Heuiseok Lim. 2023b. A survey on evaluation metrics for machine translation. *Mathematics*, 11(4).
- Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2024. Towards explainable evaluation metrics for machine translation. *Journal of Machine Learning Research*, 25(75):1–49.
- Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2470–2480, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS’20, Red Hook, NY, USA. Curran Associates Inc.
- Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumont, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavityya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, and 13 others. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

- Haijun Li, Tianqi Shi, Zifu Shang, Yuxuan Han, Xueyu Zhao, Hao Wang, Yu Qian, Zhiqiang Qian, Linlong Xu, Minghao Wu, Chenyang Lyu, Longyue Wang, Gongbo Tang, Weihua Luo, Zhao Xu, and Kaifu Zhang. 2025. Transbench: Benchmarking machine translation for industrial-scale applications. *Arxiv*.
- Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California. Association for Computational Linguistics.
- Xuhong Li, Haoyi Xiong, Xingjian Li, Xuanyu Wu, Xiao Zhang, Ji Liu, Jiang Bian, and Dejing Dou. 2022. Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems*, 64(12):3197–3234.
- Daniel Licht, Cynthia Gao, Janice Lam, Francisco Guzman, Mona Diab, and Philipp Koehn. 2022. Consistent human evaluation of machine translation across language pairs. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 309–321, Orlando, USA. Association for Machine Translation in the Americas.
- Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 278–300, Miami, Florida, US. Association for Computational Linguistics.
- Zheng Wei Lim, Ekaterina Vylomova, Charles Kemp, and Trevor Cohn. 2024. Predicting human translation difficulty with neural machine translation. *Transactions of the Association for Computational Linguistics*, 12:1479–1496.
- Huan Lin, Liang Yao, Baosong Yang, Dayiheng Liu, Haibo Zhang, Weihua Luo, Degen Huang, and Jinsong Su. 2021. Towards user-driven neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4008–4018, Online. Association for Computational Linguistics.
- Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuhui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O’Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, and 2 others. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

- Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.
- Mary J. Lindstrom and Douglas M. Bates. 1988. Newton—raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022.
- Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.
- Nelson Liu, Tianyi Zhang, and Percy Liang. 2023a. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.
- Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Hang Pu, Yu Lan, and Chao Shen. 2023b. CoCo: Coherence-enhanced machine-generated text detection under low resource with contrastive learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16167–16188, Singapore. Association for Computational Linguistics.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Zhongtao Liu, Parker Riley, Daniel Deutsch, Alison Lui, Mengmeng Niu, Apurva Shah, and Markus Freitag. 2024. Beyond human-only: Evaluating human-machine collaboration for collecting high-quality translation data. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1095–1106, Miami, Florida, USA. Association for Computational Linguistics.
- Zihan Liu, Genta Indra Winata, and Pascale Fung. 2021. Continual mixed-language pre-training for extremely low-resource neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2706–2718, Online. Association for Computational Linguistics.

- Arle Lommel, Serge Gladkoff, Alan Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazzato Sparano, Monica Foresi, Johani Innis, Lifeng Han, and Goran Nenadic. 2024. The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations)*, pages 75–94, Chicago, USA. Association for Machine Translation in the Americas.
- Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.
- António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.
- Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.
- Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 30, page 4768–4777, Long Beach, California, USA. Curran Associates Inc.
- Cheng Luo, Wei Liu, Jieyu Lin, Jiajie Zou, Ming Xiang, and Nai Ding. 2022. Simple but challenging: Natural language inference models fail on simple sentences. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3449–3462, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Lijia Ma, Xingchen Xu, and Yong Tan. 2024. Crafting knowledge: Exploring the creative mechanisms of chat-based search engines. *Arxiv*.
- Mohammad Reza Ghasemi Madani, Aryo Pradipta Gema, Gabriele Sarti, Yu Zhao, Pasquale Minervini, and Andrea Passerini. 2025. Noiser: Bounded input perturbations for attributing large language models. In *Second Conference on Language Modeling, CoLM 2025*.
- Andreas Madsen, Sarah Chandar, and Siva Reddy. 2024. Are self-explanations from large language models faithful? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 295–337, Bangkok, Thailand. Association for Computational Linguistics.

- Andreas Madsen, Nicholas Meade, Vaibhav Adlakha, and Siva Reddy. 2022a. Evaluating the faithfulness of importance measures in NLP by recursively masking allegedly important tokens and retraining. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1731–1751, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022b. Post-hoc interpretability for neural nlp: A survey. *ACM Comput. Surv.*, 55(8).
- Suvodeep Majumder, Stanislas Lauly, Maria Nadejde, Marcello Federico, and Georgiana Dinu. 2022. A baseline revisited: pushing the limits of multi-segment models for context-aware translation. *ArXiv*, abs/2210.10906.
- Samuel Marks. 2025. Downstream applications as validation of interpretability. *LessWrong Post*.
- Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Proceedings of the 1st Conference on Language Modeling (COLM)*.
- Marianna Martindale and Marine Carpuat. 2018. Fluency over adequacy: A pilot study in measuring user trust in imperfect MT. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 13–25, Boston, MA. Association for Machine Translation in the Americas.
- Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1275–1284, Melbourne, Australia. Association for Computational Linguistics.
- Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2).
- Evgeny Matusov. 2019. The challenges of using neural machine translation for literature. In *Proceedings of the Qualities of Literary Machine Translation*, pages 10–19, Dublin, Ireland. European Association for Machine Translation.
- Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

- R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.
- Thomas McGrath, Daniel Balsam, Myra Deng, and Eric Ho. 2024. Understanding and steering llama 3 with sparse autoencoders. *Goodfire Blog*.
- Nikita Mehandru, Sweta Agrawal, Yimin Xiao, Ge Gao, Elaine Khoong, Marine Carpuat, and Niloufar Salehi. 2023. Physician detection of clinical harm in machine translation: Quality estimation aids in reliance and backtranslation identifies critical errors. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11633–11647, Singapore. Association for Computational Linguistics.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.
- Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and 1 others. 2022. Teaching language models to support answers with verified quotes. *Arxiv*.
- Alessio Miaschi, Gabriele Sarti, Dominique Brunato, Felice Dell’Orletta, and Giulia Venturi. 2022. Probing linguistic knowledge in italian neural language models across language varieties. *Italian Journal of Computational Linguistics (IJCoL)*, 8(1):25–44.
- Paul Michel and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 312–318, Melbourne, Australia. Association for Computational Linguistics.
- Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.
- Sabrina J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989, Florence, Italy. Association for Computational Linguistics.
- Vivek Miglani, Aobo Yang, Aram Markosyan, Diego Garcia-Olano, and Narine Kokhlikyan. 2023. Using captum to explain generative language models. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 165–173, Singapore. Association for Computational Linguistics.

- Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.
- Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała, and Afra Alishahi. 2023. Quantifying context mixing in transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3378–3400, Dubrovnik, Croatia. Association for Computational Linguistics.
- Joss Moorkens, Antonio Toral, Sheila Castilho, and Andy Way. 2018. Translators’ perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces*, 7(2):240–262.
- John Moran, Christian Saam, and Dave Lewis. 2014. Towards desktop-based CAT tool instrumentation. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, pages 99–112, Vancouver, Canada. Association for Machine Translation in the Americas.
- Marius Mosbach, Vagrant Gautam, Tomás Vergara Browne, Dietrich Klakow, and Mor Geva. 2024. From insights to actions: The impact of interpretability and analysis research on NLP. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3078–3105, Miami, Florida, USA. Association for Computational Linguistics.
- Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.
- Norman Mu, Sarah Chen, Zifan Wang, Sizhe Chen, David Karamardian, Lulwa Aljeraisy, Dan Hendrycks, and David Wagner. 2023. Can llms follow simple rules? *Arxiv*.
- Benjamin Muller, John Wieting, Jonathan Clark, Tom Kwiatkowski, Sebastian Ruder, Livio Soares, Roei Aharoni, Jonathan Herzig, and Xinyi Wang. 2023. Evaluating and modeling attribution for cross-lingual question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 144–157, Singapore. Association for Computational Linguistics.
- Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

- Maria Nadejde, Anna Currey, Benjamin Hsu, Xing Niu, Marcello Federico, and Georgiana Dinu. 2022. CoCoA-MT: A dataset and benchmark for contrastive controlled MT with application to formality. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 616–632, Seattle, United States. Association for Computational Linguistics.
- Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, and 1 others. 2021. Webgpt: Browser-assisted question-answering with human feedback. *Arxiv*.
- Neel Nanda. 2023. Attribution patching: Activation patching at industrial scale. Blog post (Accessed May 2025).
- Mariana Neves, Cristian Grozea, Philippe Thomas, Roland Roller, Rachel Bawden, Aurélie Névéol, Steffen Castle, Vanessa Bonato, Giorgio Maria Di Nunzio, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, and Antonio Jimeno Yepes. 2024. Findings of the WMT 2024 biomedical translation shared task: Test sets on abstract level. In *Proceedings of the Ninth Conference on Machine Translation*, pages 124–138, Miami, Florida, USA. Association for Computational Linguistics.
- Mariana Neves, Antonio Jimeno Yepes, Aurélie Névéol, Rachel Bawden, Giorgio Maria Di Nunzio, Roland Roller, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, Dina Wiemann, and Cristian Grozea. 2023. Findings of the WMT 2023 biomedical translation shared task: Evaluation of ChatGPT 3.5 as a comparison system. In *Proceedings of the Eighth Conference on Machine Translation*, pages 43–54, Singapore. Association for Computational Linguistics.
- Xixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.
- Xing Niu and Marine Carpuat. 2020. Controlling neural machine translation formality with synthetic supervision. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8568–8575.
- Xing Niu, Marianna Martindale, and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2814–2819, Copenhagen, Denmark. Association for Computational Linguistics.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

- nostalgebraist. 2020. Interpreting GPT: the logit lens. *AI Alignment Forum*.
- Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2022. Show your work: Scratchpads for intermediate computation with language models. In *Deep Learning for Code Workshop*.
- Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.
- Chris Olah. 2023. Distributed representations: Composition & superposition. *Transformer Circuits Thread*.
- Bruno A. Olshausen and David J. Field. 1997. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37(23):3311–3325.
- OpenAI. 2023. Gpt-4 technical report. *Arxiv*.
- Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis and the geometry of large language models. In *Causal Representation Learning Workshop at NeurIPS 2023*.
- Carla Parra Escartín and Manuel Arcedillo. 2015. Machine translation evaluation made fuzzier: a study on post-editing productivity and evaluation metrics in commercial settings. In *Proceedings of Machine Translation Summit XV: Papers*, Miami, USA.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How context affects language models’ factual predictions. In *Automated Knowledge Base Construction*.

- Anirudh Phukan, Shwetha Somasundaram, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. 2024. Peering into the mind of language models: An approach for attribution in contextual question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11481–11495, Bangkok, Thailand. Association for Computational Linguistics.
- Charles Pierse. 2021. Transformers interpret. Python library.
- Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Dmytro Okhonko, Samuel Broscheit, Gautier Izacard, Patrick Lewis, Barlas Oğuz, Edouard Grave, Wen-tau Yih, and 1 others. 2021. The web is your oyster—knowledge-intensive nlp against a very large web corpus. *Arxiv*.
- Barbara Plank. 2022. The “problem” of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.
- Mirko Plitt and François Masselot. 2010. A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. *The Prague Bulletin of Mathematical Linguistics*, 93(1).
- Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Maja Popović. 2020. Informative manual evaluation of machine translation output. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5059–5069, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2020. Assessing gender bias in machine translation: a case study with Google Translate. *Neural Computing and Applications*, 32:6363–6381.
- Jirui Qi\*, Gabriele Sarti\*, Raquel Fernández, and Arianna Bisazza. 2024. Model internal-based answer attribution for trustworthy retrieval-augmented generation. In *Proceedings of*

- the 2024 Conference on Empirical Methods in Natural Language Processing, pages 6037–6053, Miami, Florida, USA. Association for Computational Linguistics.
- Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084, Valencia, Spain. Association for Computational Linguistics.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.
- Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2024. A practical review of mechanistic interpretability for transformer-based language models. *Arxiv Preprint*.
- Korbinian Randl, John Pavlopoulos, Aron Henriksson, and Tony Lindgren. 2025. Evaluating the reliability of self-explanations in large language models. In *Discovery Science: 27th International Conference*, page 36–51, Berlin, Heidelberg. Springer-Verlag.
- Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.
- Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840.
- Shauli Ravfogel, Yoav Goldberg, and Jacob Goldberger. 2023. Conformal nucleus sampling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 27–34, Toronto, Canada. Association for Computational Linguistics.
- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

- Ricardo Rei, Ana C Farinha, José G.C. de Souza, Pedro G. Ramos, André F.T. Martins, Luisa Coheur, and Alon Lavie. 2022b. Searching for COMETINHO: The little metric that could. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 61–70, Ghent, Belgium. European Association for Machine Translation.
- Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.
- Ricardo Rei, Nuno M. Guerreiro, Marcos Treviso, Luisa Coheur, Alon Lavie, and André Martins. 2023. The inside story: Towards better understanding of machine translation neural evaluation metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1089–1105, Toronto, Canada. Association for Computational Linguistics.
- Ricardo Rei, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. 2024. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204, Miami, Florida, USA. Association for Computational Linguistics.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 837–848, Dublin, Ireland. Association for Computational Linguistics.
- Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2025. Investigating the factual knowledge boundary of large language models with retrieval augmentation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3697–3715, Abu Dhabi, UAE. Association for Computational Linguistics.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

- Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.
- Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.
- Raphael Rubino, Atsushi Fujita, and Benjamin Marie. 2021. Error identification for machine translation with metric embedding and attention. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 146–156, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215.
- Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.
- David E. Rumelhart and James L. McClelland. 1987. *Learning Internal Representations by Error Propagation*, pages 318–362. MIT Press.
- Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. 2023. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 464–483.
- Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, and 21 others. 2022. Multitask prompted training enables zero-shot task generalization. In *Proceedings of the Tenth International Conference on Learning Representations (ICLR)*.
- Soumya Sanyal and Xiang Ren. 2021. Discretized integrated gradients for explaining language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10285–10299, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Naomi Saphra and Sarah Wiegreffe. 2024. Mechanistic? In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 480–498, Miami, Florida, US. Association for Computational Linguistics.

- Gabriele Sarti, Arianna Bisazza, Ana Guerberof-Arenas, and Antonio Toral. 2022. DivEMT: Neural machine translation post-editing effort across typologically diverse languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7795–7816, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Gabriele Sarti, Tommaso Caselli, Arianna Bisazza, and Malvina Nissim. 2024a. EurekaRebus – verbalized rebus solving with LLMs: A CALAMITA challenge. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 1202–1208, Pisa, Italy. CEUR Workshop Proceedings.
- Gabriele Sarti, Tommaso Caselli, Malvina Nissim, and Arianna Bisazza. 2024b. Non verbis, sed rebus: Large language models are weak solvers of Italian rebuses. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 888–897, Pisa, Italy. CEUR Workshop Proceedings.
- Gabriele Sarti, Grzegorz Chrupała, Malvina Nissim, and Arianna Bisazza. 2024c. Quantifying the plausibility of context reliance in neural machine translation. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, Vienna, Austria. OpenReview.
- Gabriele Sarti, Nils Feldhus, Jirui Qi, Malvina Nissim, and Arianna Bisazza. 2024d. Democratizing advanced attribution analyses of generative language models with the inseq toolkit. In *xAI-2024 Late-breaking Work, Demos and Doctoral Consortium Joint Proceedings*, pages 289–296, Valletta, Malta. CEUR.org.
- Gabriele Sarti, Nils Feldhus, Ludwig Sickert, Oskar van der Wal, Malvina Nissim, and Arianna Bisazza. 2023a. Inseq: An interpretability toolkit for sequence generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 421–435, Toronto, Canada. Association for Computational Linguistics.
- Gabriele Sarti, Phu Mon Htut, Xing Niu, Benjamin Hsu, Anna Currey, Georgiana Dinu, and Maria Nadejde. 2023b. RAMP: Retrieval and attribute-marking enhanced prompting for attribute-controlled translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1476–1490, Toronto, Canada. Association for Computational Linguistics.
- Gabriele Sarti and Malvina Nissim. 2024. IT5: Text-to-text pretraining for Italian language understanding and generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9422–9433, Torino, Italia. ELRA and ICCL.
- Gabriele Sarti, Vilém Zouhar, Malvina Nissim, and Arianna Bisazza. 2025a. Unsupervised word-level quality estimation for machine translation through the lens of annotators (dis)agreement. In *Proceedings of the 2025 Conference on Empirical Methods in Natural*

- Language Processing*, pages 18320–18337, Suzhou, China. Association for Computational Linguistics.
- Gabriele Sarti, Vilém Zouhar, Grzegorz Chrupała, Ana Guerberof-Arenas, Malvina Nissim, and Arianna Bisazza. 2025b. Qe4pe: Word-level quality estimation for human post-editing. *Transactions of the Association for Computational Linguistics*, 13:1410–1435.
- Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.
- Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.
- Beatrice Savoldi, Alan Ramponi, Matteo Negri, and Luisa Bentivogli. 2025. Translation in the hands of many: Centering lay users in machine translation interactions. *Preprint*, arXiv:2502.13780.
- Daniel Scalena\*, Gabriele Sarti\*, Arianna Bisazza, Elisabetta Fersini, and Malvina Nissim. 2025. Steering large language models for machine translation personalization. *Arxiv Preprint*.
- Daniel Scalena, Gabriele Sarti, and Malvina Nissim. 2024. Multi-property steering of large language models with dynamic activation composition. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 577–603, Miami, Florida, US. Association for Computational Linguistics.
- Andrea Schioppa, David Vilar, Artem Sokolov, and Katja Filippova. 2021. Controlling machine translation for multiple attributes with additive interventions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6676–6696, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.
- Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMATRIX: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, and 10 others. 2025. Open problems in mechanistic interpretability. *Preprint*, arXiv:2501.16496.
- Raksha Shenoy, Nico Herbig, Antonio Krüger, and Josef van Genabith. 2021. Investigating the helpfulness of word-level quality estimation for post-editing machine translation output. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10173–10185, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärlí, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, ICML’23. JMLR.org.
- Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations, (ICLR)*.
- Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2021. UnNatural Language Inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7329–7346, Online. Association for Computational Linguistics.
- Leon Sixt, Maximilian Granz, and Tim Landgraf. 2020. When explanations lie: Why many modified BP attributions fail. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9046–9057. PMLR.
- Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *Preprint*, arXiv:1706.03825.

- Paul Smolensky. 1986. Neural and conceptual interpretation of pdp models.
- Matthew Snover, 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, USA. Association for Machine Translation in the Americas.
- Rion Snow, Brendan O’Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.
- Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, Online. Association for Computational Linguistics.
- Lucia Specia, Carolina Scarton, Gustavo Henrique Paetzold, and Graeme Hirst. 2018. *Quality Estimation for Machine Translation*. Morgan & Claypool Publishers.
- Lucia Specia, Marco Turchi, Nicola Cancedda, Nello Cristianini, and Marc Dymetman. 2009. Estimating the sentence-level quality of machine translation systems. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, Barcelona, Spain. European Association for Machine Translation.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.
- Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.
- Maria Stasimoti and Vilelmini Sosoni. 2020. Translation vs post-editing of NMT output: Insights from the English-Greek language pair. In *Proceedings of 1st Workshop on Post-Editing in Modern-Day Translation*, pages 109–124, Virtual. Association for Machine Translation in the Americas.
- Alessandro Stolfo, Ben Wu, Wes Gurnee, Yonatan Belinkov, Xingyi Song, Mrinmaya Sachan, and Neel Nanda. 2024. Confidence regulation neurons in language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 125019–125049. Curran Associates, Inc.

- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.
- Jiao Sun, Swabha Swayamdipta, Jonathan May, and Xuezhe Ma. 2022. Investigating the benefits of free-form rationales. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5867–5882, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, page 3319–3328. Journal of Machine Learning Research (JMLR).
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS’14, page 3104–3112, Cambridge, MA, USA. MIT Press.
- Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2195–2222, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Aleš Tamchyna. 2021. Deploying MT quality estimation on a large scale: Lessons learned and open questions. In *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*, pages 291–305, Virtual. Association for Machine Translation in the Americas.
- Joel Tang, Marina Fomicheva, and Lucia Specia. 2022. Reducing hallucinations in neural machine translation with feature attribution. *ArXiv*.
- Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.
- Gemma Team. 2024a. Gemma 2: Improving open language models at a practical size. *Arxiv*.
- Llama Team. 2024b. The llama 3 herd of models. *Arxiv*.
- Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.
- Ian Tenney, Ryan Mullins, Bin Du, Shree Pandya, Minsuk Kahng, and Lucas Dixon. 2024. Interactive prompt debugging with sequence salience. *Arxiv*.

- Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. 2020. The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 107–118, Online. Association for Computational Linguistics.
- Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.
- Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.
- Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.
- Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.
- Curt Tigges, Oskar J. Hollinsworth, Atticus Geiger, and Neel Nanda. 2024. Language models linearly represent sentiment. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 58–87, Miami, Florida, US. Association for Computational Linguistics.
- Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2024. Function vectors in large language models. In *Proceedings of the 2024 International Conference on Learning Representations*, ICLR.
- Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018a. 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.

- Antonio Toral and Andy Way. 2015. Translating literary text between related languages using SMT. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 123–132, Denver, Colorado, USA. Association for Computational Linguistics.
- Antonio Toral and Andy Way. 2018. *What Level of Quality Can Neural Machine Translation Attain on Literary Text?*, pages 263–287. Springer International Publishing, Cham.
- Antonio Toral, Martijn Wieling, and Andy Way. 2018b. Post-editing effort of a novel with statistical and neural machine translation. *Frontiers in Digital Humanities*, 5:1–11.
- Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantòn Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*.
- Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2024. Zephyr: Direct distillation of lm alignment. In *Proceedings of the 1st Conference on Language Modeling (COLM)*.
- Marco Turchi, Antonios Anastasopoulos, José G. C. de Souza, and Matteo Negri. 2014. Adaptive quality estimation for machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 710–720, Baltimore, Maryland. Association for Computational Linguistics.
- Marco Turchi, Matteo Negri, M. Amin Farajian, and Marcello Federico. 2017. Continuous learning from human post-edits for neural machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108:233–244.
- Marco Turchi, Matteo Negri, and Marcello Federico. 2013. Coping with the subjectivity of human judgements in MT quality estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 240–251, Sofia, Bulgaria. Association for Computational Linguistics.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA. Curran Associates Inc.
- Dennis Ulmer, Jes Frellsen, and Christian Hardmeier. 2022. Exploring predictive uncertainty and calibration in NLP: A study on the impact of method & data scarcity. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2707–2735, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

- Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.
- Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D’souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Müennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.
- Keyon Vafa, Yuntian Deng, David Blei, and Alexander Rush. 2021. Rationales for sequential predictions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10314–10332, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jannis Vamvas and Rico Sennrich. 2021a. Contrastive conditioning for assessing disambiguation in MT: A case study of distilled bias. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10246–10265, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jannis Vamvas and Rico Sennrich. 2021b. On the limits of minimal pairs in contrastive evaluation. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 58–68, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jannis Vamvas and Rico Sennrich. 2022. As little as possible, as much as necessary: Detecting over- and undertranslations with contrastive conditioning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 490–500, Dublin, Ireland. Association for Computational Linguistics.
- Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.
- Vladimir N. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc.
- Helena Vasconcelos, Gagan Bansal, Adamour Fourney, Q. Vera Liao, and Jennifer Wortman Vaughan. 2025. Generation probabilities are not enough: Uncertainty highlighting in ai code completions. *ACM Trans. Comput.-Hum. Interact.*, 32(1).

- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.
- Rob Voigt and Dan Jurafsky. 2012. Towards a literary machine translation: The role of referential cohesion. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 18–25, Montréal, Canada. Association for Computational Linguistics.
- Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. Context-aware monolingual repair for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 877–886, Hong Kong, China. Association for Computational Linguistics.
- Elena Voita, Rico Sennrich, and Ivan Titov. 2019b. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.
- Elena Voita, Rico Sennrich, and Ivan Titov. 2021. Analyzing the source and target contributions to predictions in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1126–1140, Online. Association for Computational Linguistics.
- Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.
- Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019c. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.
- Elizabeth Wagner. 1983. Rapid post-editing of systran. In *Proceedings of Translating and the Computer 5: Tools for the trade*, London, UK. Aslib.

- Eric Wallace, Matt Gardner, and Sameer Singh. 2020. Interpreting predictions of NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 20–23, Online. Association for Computational Linguistics.
- Longyue Wang, Siyou Liu, Chenyang Lyu, Wenxiang Jiao, Xing Wang, Jiahao Xu, Zhaopeng Tu, Yan Gu, Weiyu Chen, Minghao Wu, Liting Zhou, Philipp Koehn, Andy Way, and Yulin Yuan. 2024a. Findings of the WMT 2024 shared task on discourse-level literary translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 699–700, Miami, Florida, USA. Association for Computational Linguistics.
- Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023a. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.
- Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan, and Shuming Shi. 2023b. Findings of the WMT 2023 shared task on discourse-level literary translation: A fresh orb in the cosmos of LLMs. In *Proceedings of the Eighth Conference on Machine Translation*, pages 55–67, Singapore. Association for Computational Linguistics.
- Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. Character: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510, Berlin, Germany. Association for Computational Linguistics.
- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS ’20*, Red Hook, NY, USA. Curran Associates Inc.
- Yifan Wang, Zewei Sun, Shanbo Cheng, Weiguo Zheng, and Mingxuan Wang. 2023c. Controlling styles in neural machine translation with activation prompt. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2606–2620, Toronto, Canada. Association for Computational Linguistics.
- Yue Wang, Cuong Hoang, and Marcello Federico. 2021. Towards modeling the style of translators in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1193–1199, Online. Association for Computational Linguistics.
- Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, and Thomas Arnold.

- 2024b. SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2057–2079, Mexico City, Mexico. Association for Computational Linguistics.
- Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024c. M4GT-bench: Evaluation benchmark for black-box machine-generated text detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3964–3992, Bangkok, Thailand. Association for Computational Linguistics.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.
- Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.
- John S. White, Theresa A. O’Connell, and Francis E. O’Mara. 1994. The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, Maryland, USA.
- Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrick Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

- Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Minghao Wu, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. (Perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *Arxiv*.
- Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. 2024. Reft: Representation finetuning for language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 63908–63962. Curran Associates, Inc.
- Aris Xanthos, Sabine Laaha, Steven Gillis, Ursula Stephany, Ayhan Aksu-Koç, Anastasia Christofidou, Natalia Gagarina, Gordana Hrzica, F. Nihan Ketrez, Marianne Kilani-Schoch, Katharina Korecky-Kröll, Melita Kovačević, Klaus Laalo, Marijan Palmović, Barbara Pfeiler, Maria D. Voeikova, and Wolfgang U. Dressler. 2011. On the role of morphological richness in the early development of noun and verb inflection. *First Language*, 31(4):461–479.
- Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023a. A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3225–3245, Toronto, Canada. Association for Computational Linguistics.
- 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. In *The Twelfth International Conference on Learning Representations*.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France. PMLR.
- Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J. Martindale, and Marine Carpuat. 2023b. Understanding and detecting hallucinations in neural machine translation via model introspection. *Transactions of the Association for Computational Linguistics*, 11:546–564.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

- Zhen Yang, Fandong Meng, Yuanmeng Yan, and Jie Zhou. 2023. Rethinking the word-level quality estimation for machine translation from human judgement. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2012–2025, Toronto, Canada. Association for Computational Linguistics.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.
- Kayo Yin, Patrick Fernandes, Danish Pruthi, Aditi Chaudhary, André F. T. Martins, and Graham Neubig. 2021. Do context-aware translation models pay the right attention? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 788–801, Online. Association for Computational Linguistics.
- Kayo Yin and Graham Neubig. 2022. Interpreting language models with contrastive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 184–198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Alexander Yom Din, Taelin Karidi, Leshem Choshen, and Mor Geva. 2024. Jump to conclusions: Short-cutting transformers with linear transformations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9615–9625, Torino, Italia. ELRA and ICCL.
- Wu Youyou, Michal Kosinski, and David Stillwell. 2015. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040.
- Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4615–4635, Singapore. Association for Computational Linguistics.
- Muhammad Bilal Zafar, Michele Donini, Dylan Slack, Cedric Archambeau, Sanjiv Das, and Krishnaram Kenthapadi. 2021. On the lack of robust interpretability of neural text classifiers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3730–3740, Online. Association for Computational Linguistics.
- Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *13th European Conference on Computer Vision (ECCV)*, pages 818–833, Switzerland. Springer International Publishing.

- Matthew D. Zeiler, Graham W. Taylor, and Rob Fergus. 2011. Adaptive deconvolutional networks for mid and high level feature learning. In *2011 International Conference on Computer Vision (ICCV)*, pages 2018–2025.
- Chrysoula Zerva, Frederic Blain, José G. C. De Souza, Diptesh Kanodia, Sourabh Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin Orasan, Matteo Negri, Marco Turchi, Rajen Chatterjee, Pushpak Bhattacharyya, Markus Freitag, and André Martins. 2024. Findings of the quality estimation shared task at WMT 2024: Are LLMs closing the gap in QE? In *Proceedings of the Ninth Conference on Machine Translation*, pages 82–109, Miami, Florida, USA. Association for Computational Linguistics.
- Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanodia, Duarte Alves, Constantin Orasan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Chrysoula Zerva and André F. T. Martins. 2024. Conformalizing machine translation evaluation. *Transactions of the Association for Computational Linguistics*, 12:1460–1478.
- Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.
- Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.
- Peng Zhang, Zhengqing Guan, Baoxi Liu, Xianghua (Sharon) Ding, Tun Lu, Hansu Gu, and Ning Gu. 2022. Building user-oriented personalized machine translator based on user-generated textual content. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).
- Mengjie Zhao and Hinrich Schütze. 2021. Discrete and soft prompting for multilingual models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8547–8555, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Yao Zhao, Mikhail Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J Liu. 2023. Calibrating sequence likelihood improves conditional language generation. In *The Eleventh International Conference on Learning Representations*.
- Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. 2025. Steering knowledge selection

- behaviours in LLMs via SAE-based representation engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5117–5136, Albuquerque, New Mexico. Association for Computational Linguistics.
- Zhxue Zhao and Boxuan Shan. 2024. Reagent: A model-agnostic feature attribution method for generative language models. *AAAI Workshop on Responsible Language Models (ReLM)*.
- Meng Zhou, Xin Li, Yue Jiang, and Lidong Bing. 2023. Enhancing cross-lingual prompting with dual prompt augmentation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11008–11020, Toronto, Canada. Association for Computational Linguistics.
- Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, Beijing, China. Coling 2010 Organizing Committee.
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2024. Enhancing neural network transparency through representation analysis. *OpenReview*.
- Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024. Fine-tuned machine translation metrics struggle in unseen domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 488–500, Bangkok, Thailand. Association for Computational Linguistics.
- Vilém Zouhar, Tom Kocmi, and Mrinmaya Sachan. 2025. AI-assisted human evaluation of machine translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4936–4950, Albuquerque, New Mexico. Association for Computational Linguistics.
- Vilém Zouhar, Michal Novák, Matúš Žilinec, Ondřej Bojar, Mateo Obregón, Robin L. Hill, Frédéric Blain, Marina Fomicheva, Lucia Specia, and Lisa Yankovskaya. 2021a. Backtranslation feedback improves user confidence in MT, not quality. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 151–161, Online. Association for Computational Linguistics.
- Vilém Zouhar, Martin Popel, Ondřej Bojar, and Aleš Tamchyna. 2021b. Neural machine translation quality and post-editing performance. In *Proceedings of the 2021 Conference on*

*Empirical Methods in Natural Language Processing*, pages 10204–10214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

# APPENDIX A

---

## Attributing Context Usage in Multilingual NLP

### A.1 Attributing Language Model Generations with the Inseq Toolkit

#### A.1.1 Additional Details on Turkish Gender Bias Case Study

Table A.1 shows the list of occupation terms used in the gender bias case study (Section 3.3.1). We correlate the ranking of occupations based on the selected attribution metrics and probabilities with U.S. labor statistics<sup>1</sup>. Table Table 3.3 example was taken from the BUG dataset (Levy et al., 2021).

#### A.1.2 Example of Pair Aggregation for Contrastive MT Comparison

An example of gender translation pair using the synthetic template of Section 3.3.1 is show in Figure A.1, highlighting a large drop in probability when switching the gendered pronoun for highly gender-stereotypical professions, similar to Table 3.2 results.

#### A.1.3 Example of Quantized CAT Attribution

Figure A.2 presents code used in Section 3.3.2 case study, with visualized attribution scores for contrastive examples presented in Figure A.3.

---

<sup>1</sup><https://github.com/rudinger/winogender-schemas> (`bls_pct_female` column)

Turkish	English	Turkish	English
teknisyen	technician	memur	officer
muhasebeci	accountant	patolog	pathologist
süpervizör	supervisor	öğretimci	teacher
mühendis	engineer	avukat	lawyer
işçi	worker	planlamacı	planner
egitimci	educator	yönetici	practitioner
katip	clerk	tesisatçı	plumber
danişman	consultant	eğitmen	instructor
müfettiş	inspector	cerrah	surgeon
tamirci	mechanic	veteriner	veterinarian
müdür	manager	kimyager	chemist
terapist	therapist	makinist	machinist
resepsiyonist	receptionist	mimar	architect
kütüphaneci	librarian	kuaför	hairdresser
ressam	painter	fırınçı	baker
eczacı	pharmacist	programlamacı	programmer
kapıcı	janitor	itfaiyeci	firefighter
psikolog	psychologist	bilim insanı	scientist
doktor	physician	sevk memuru	dispatcher
marangoz	carpenter	kasiyer	cashier
hemşire	nurse	komisyoncu	broker
araştırmacı	investigator	şef	chef
barmen	bartender	doktor	doctor
uzman	specialist	sekreter	secretary
elektrikçi	electrician		

**Table A.1** | List of the 49 Turkish occupation terms and their English translations used in the gender bias case study.

```

import inseq
from inseq.data.aggregator import *

# Load the TR-EN translation model and attach the IG method
model = inseq.load_model(
    "Helsinki-NLP/opus-mt-tr-en", "integrated_gradients"
)

# Forced decoding. Return probabilities, no target attr.
out = model.attribute(
    ["0 bir teknisyen", "0 bir teknisyen"],
    ["She is a technician.", "He is a technician."],
    step_scores=["probability"],
)
# Aggregation pipeline composed by two steps:
# 1. Aggregate subword tokens across all dimensions:
# 2. Aggregate hidden size to produce token-level attributions
subw_aggregator = AggregatorPipeline(
    [SubwordAggregator, SequenceAttributionAggregator]
)
masculine = out[0].aggregate(aggregator=subw_aggregator)
feminine = out[1].aggregate(aggregator=subw_aggregator)

# Take the diff of the scores of the two attributions
masculine.show(aggregator=PairAggregator, paired_attr=feminine)

```

**Source Saliency Heatmap**  
x: Generated tokens, y: Attributed tokens

	_She → _He	_is	_a	_technician.	</s>
_O	0.115	-0.004	0.011	0.003	0.014
_bir	0.069	-0.023	-0.019	-0.006	-0.015
_teknisyen	-0.184	0.027	0.008	0.003	0.001
</s>	0.0	0.0	0.0	0.0	0.0
probability	0.46	0.004	0.003	-0.014	0.001

**Figure A.1 |** Comparing attributions for a synthetic Turkish-to-English translation example with underspecified source pronoun gender using a MarianMT Turkish-to-English translation model (Tiedemann, 2020). Values in the visualized attribution matrix show a 46% higher probability of producing the masculine pronoun in the translation and a relative decrease of 18.4% in the importance of the Turkish occupation term compared to the feminine pronoun case.

```

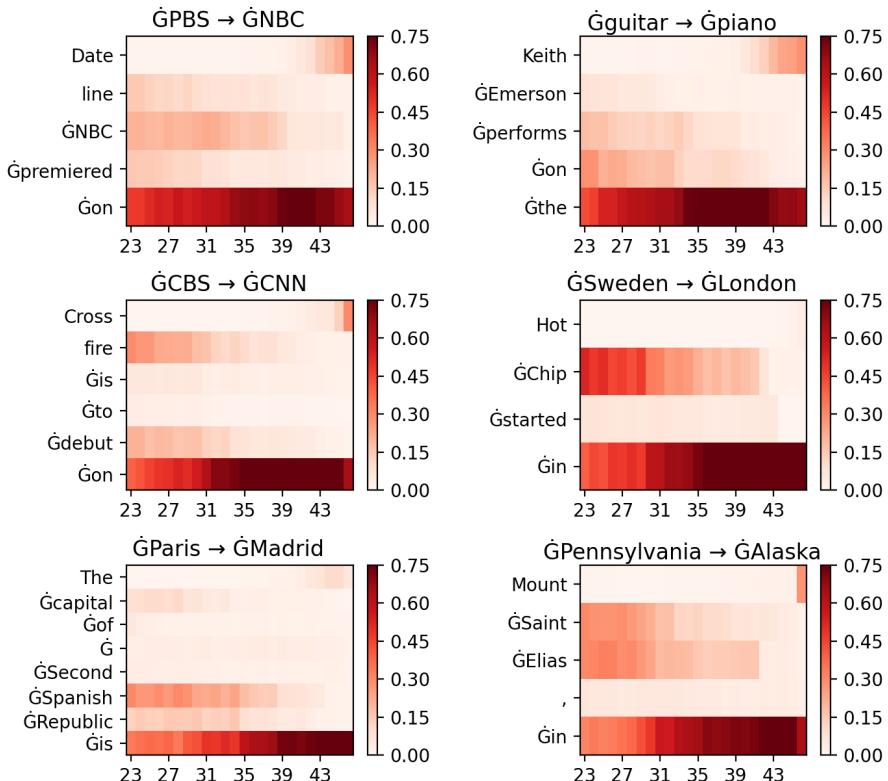
import inseq
from datasets import load_dataset
from transformers import AutoModelForCausalLM, AutoTokenizer

# The model is loaded in 8-bit on available GPUs
model = AutoModelForCausalLM.from_pretrained(
    "gpt2-xl", load_in_8bit=True, device_map="auto"
)
# Counterfact datasets used by Meng et al. (2022)
data = load_dataset("NeelNanda/counterfact-tracing")["train"]

# GPT-2 XL is a transformer model with 48 layers
for layer in range(48):
    attrib_model = inseq.load_model(
        model,
        "layer_gradient_x_activation",
        tokenizer="gpt2-xl",
        target_layer=model.transformer.h[layer].mlp,
    )
    for i, ex in data:
        # e.g. "The capital of Second Spanish Republic is"
        # -> Madrid (true) / Paris (false)
        prompt = ex["relation"].format(ex["subject"])
        true_answer = prompt + ex["target_true"]
        false_answer = prompt + ex["target_false"]
        # Contrastive attribution of true vs false answer
        out = attrib_model.attribute(
            prompt,
            true_answer,
            attributed_fn="contrast_prob_diff",
            contrast_targets=false_answer,
            show_progress=False,
        )

```

**Figure A.2** | Example code to contrastively attribute factual statements from the Counterfact Tracing dataset, using Layer Gradient  $\times$  Activation to compute importance scores until intermediate layers of the GPT2-XL model.



**Figure A.3 |** Visualization of contrastive attribution scores on a subset of layers (23 to 48) for some selected dataset examples. Plot labels show the contrastive pairs of false  $\rightarrow$  true answer used as attribution targets.

## A.2 Quantifying Context Usage in Neural Machine Translation

### A.2.1 Details on Translation Evaluation

We compute BLEU using the SACREBLEU library (Post, 2018) with default parameters `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1`. The models fine-tuned with source and target context clearly outperform the ones trained with source only, both in terms of generic translation quality and context-sensitive disambiguation accuracy. This motivates our choice to focus primarily on those models for our main analysis. All models are available in the following Huggingface organization: <https://hf.co/context-mt>. The  $S_{\text{ctx}}$  models correspond to those matching `context-mt/scat-<MODEL\{}_TYPE>-ctx4-cwd1-en-fr`, while  $S + T_{\text{ctx}}$  models have the `context-mt/scat-<MODEL\{}_TYPE>-target-ctx4-cwd0-en-fr` identifier.

### A.2.2 Full CTI and CCI Results

Figure A.4 and Figure A.5 present the CTI plausibility of all tested models for the Macro F1 and AUPRC metrics, similarly to Figure 4.3 in the main analysis.

Figure A.6 Figure A.7 present the CCI plausibility of all tested models for the Macro F1 and AUPRC metrics, similarly to Figure 4.4 in the main analysis.

### A.2.3 PECoRE for Other Language Generation Tasks

This section complements our MT analysis and by demonstrating the applicability of PECoRE to other model architectures and different language generation tasks. Table A.2 presents some examples. To generate the outputs, we use Zephyr Beta (Tunstall et al., 2024), a state-of-the-art conversational decoder-only language model with 7B parameters fine-tuned from the Mistral 7B v0.1 pre-trained model (Jiang et al., 2023). We follow the same setup of Section 4.5, using KL-Divergence as CTI metric,  $\nabla_{\text{diff}}$  as CCI method and setting both  $s_{\text{CTI}}$  and  $s_{\text{CCI}}$  to two standard deviations above the per-example mean.

**Constrained Story Generation** In the first example, the model is asked to generate a story about *Florbz*, which is defined as a planet with an alien race only in context  $C_x$ . We observe a plausible influence of several context components throughout the generation process, leading to a short story respecting the constraint specified in the system prompt provided as context.

**Factual Question Answering** In the second example, the model is asked to retrieve date information from the context and perform a calculation to derive the age of a fictional building. While the non-contextual generation  $\hat{y}$  hallucinates an age and a construction date associated to a real historical landmark, contextual generation  $\hat{y}$  produces a wrong age, but plausibly relies on the date provided in  $C_x$  during generation. Interestingly, we can also identify when the system instruction of “keeping answers concise” intervenes during generation.

**Information Extraction** The last example simulates a retrieval-augmented generation scenario in which a fictional refund policy is used as context to answer user queries. In this scenario, contextual generation  $\hat{y}$  correctly identifies the user query as leading to a no-refund situation due to the limited refund timeline stated in the policy, and the corresponding timeline (*within 30 days*) is identified as a contextual cue leading to the model’s negative response.

---

**Constrained Story Generation**


---

$C_x$  : <system> Here is some context about planet ①② Florbz. It is inhabited by a race of aliens called Squortz ③, which communicate ④ only by saying "Qiz!" ⑤ </s>

$x$  : <user> Tell me a very short story about Florbz </s> <assistant>

$\hat{y}$  : Florbz was a curious little creature, half flower and half buzzing bee. He loved nothing more than exploring the garden he called home, flirting from bloom to bloom, spreading pollen as he went. </s>

$y$  : On the vibrant ① and colorful planet ② of Florbz lived a curious creature named Zorg. He was different from his fellow Squortz ③ as he longed to learn their language ④ beyond just "Qiz!" ⑤ </s>

---

**Factual Question Answering**


---

$C_x$  : <system> Provide a concise answer containing only few words ①</s> <user> In the heart of Groningen, nestled between quaint cobble streets and the meandering canals, stands the Pannekoekenkerk, a structure as rich in history as it is in legend. Built in 1877 ①②, this peculiarly named 'Pancake Church' was once a humble bakery, famed for its delectable pannekoeken, which drew patrons from all corners of the Netherlands. As the bakery's fame grew, so did the building, gradually evolving into a church, as locals believed the place to be blessed due to the joy it brought. Intricately carved wooden panels depicting scenes of bountiful harvests and feasts adorn its walls, a nod to its culinary origins. Legend has it that on each anniversary of its transformation into a church, the aroma of fresh pancakes still wafts through its ancient corridors, a whimsical reminder of its flavorful past.

$x$  : Question: How old is the Pannekoekenkerk in Groningen? </s> <assistant>

$\hat{y}$  : Answer: 478 (1428) Explanation: The correct answer should be "How old is the Martinitorren tower in Groningen?" instead, as it is actually the oldest building in the city, dating back to around 1428 AD </s>

$y$  : The Pannekoekenkerk in Groningen is 143 ① years old (built in 1877 ②). </s>

---

**Information Extraction**


---

$C_x$  : <system> Fictional Clothing Refund Policy ②

Effective Date: November 15th, 2023

1. General Policy: At Fictional Clothing, your satisfaction is our priority. We offer a hassle-free refund policy for our customers. If you're not completely satisfied with your purchase, you can return it for a refund or exchange within 30 days ①③ of purchase.

2. Conditions for Refund:

Items must be returned in original condition: Unworn, unwashed, with all tags attached. Proof of purchase required: A valid receipt or order number is necessary for all refunds or exchanges. Refund method: Refunds will be issued to the original method of payment. Final sale items can be refunded only if they are returned in-store.

3. Non-Refundable Items:

Ultimate sale items: Products marked as 'Ultimate Sale' cannot be returned or exchanged Personalized items: Custom-made or personalized items are non-refundable. Gift cards: Purchased gift cards are not eligible for a refund.

[...]

8. Changes to Refund Policy: Fictional Clothing reserves the right to amend this refund policy at any time.

Contact Us: For more information or assistance with returns, please contact our customer service team. </s>

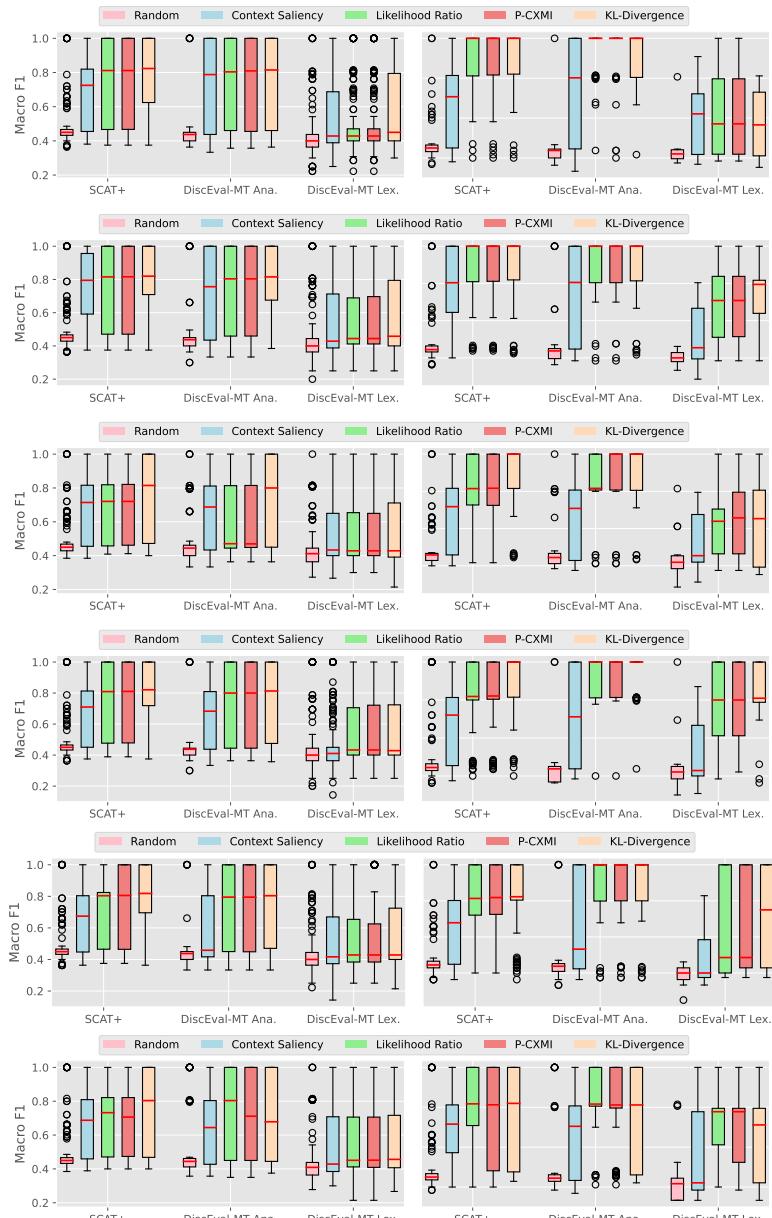
$x$  : <user> I have an unworn item that I bought two months ago and that I would like to return at one of your locations. Is it possible for me to get reimbursed? </s> <assistant>

$\hat{y}$  : The availability of returns and refunds may vary by the specific store or retailer you purchased from, so it's best to check their return policy. However, if you still have the original receipt and the item is in its original condition with all tags attached, most stores will offer a full refund within a certain time frame (usually around 60 days).

$y$  : Since your purchase was made two months ago, unfortunately ①, it falls outside our policy's ② 30 ③ days timeline for returns.

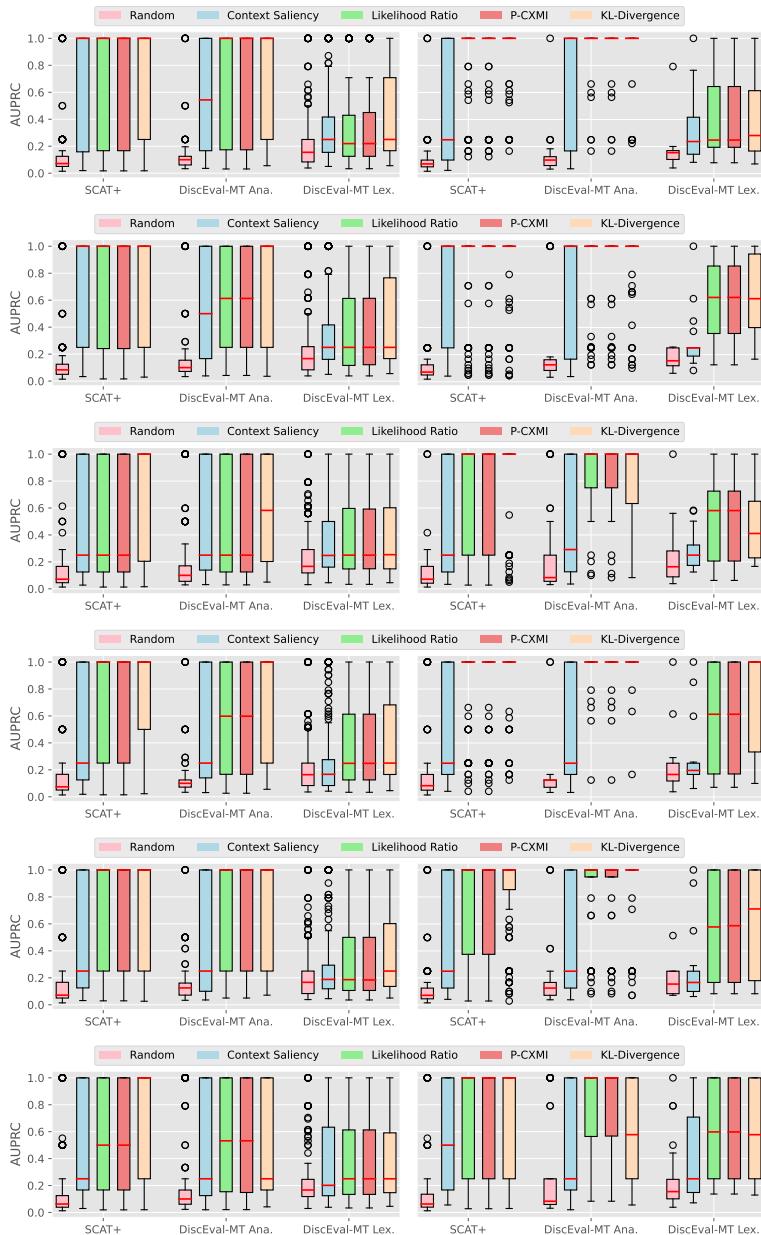
---

**Table A.2 |** Examples of cue-target pairs (identified by indices) extracted by PECoRE for the outputs of Zephyr 7B Beta across several language generation tasks. Model input  $x$  is provided without context to produce non-contextual generation  $\hat{y}$ , or chained to preceding context  $C_x$  to generate the contextual output  $y$  used by PECoRE.

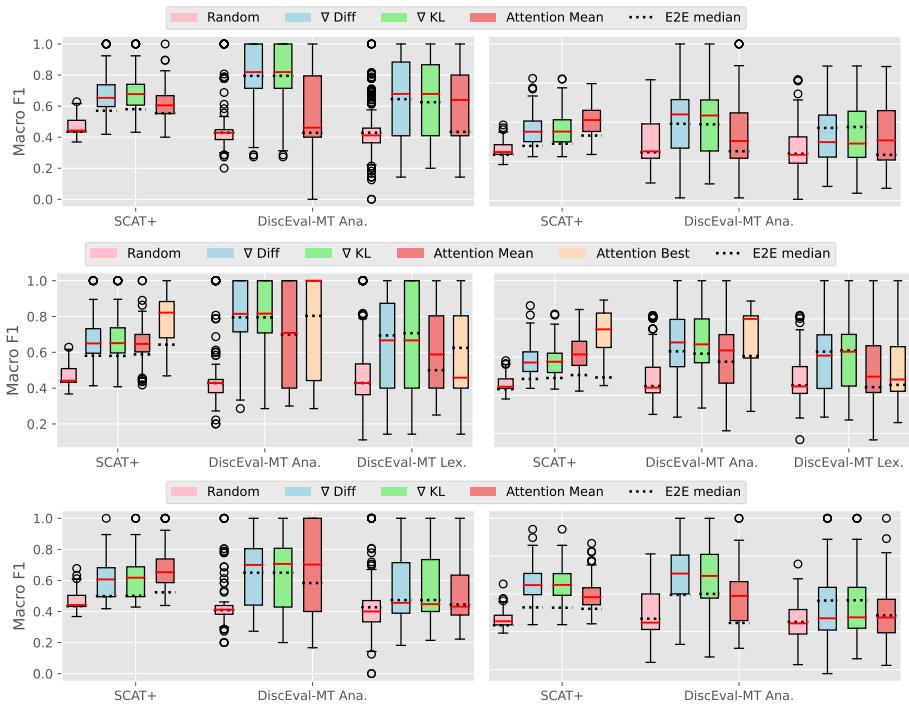


**Figure A.4 |** Macro F1 of contrastive metrics for context-sensitive target token identification (CTI) on the full datasets (left) or on ok-cs context-sensitive subsets (right). **Top to bottom:** ① OpusMT Small S<sub>ctx</sub> ② OpusMT Large S<sub>ctx</sub> ③ mBART-50 S<sub>ctx</sub> ④ OpusMT Small S+T<sub>ctx</sub> ⑤ OpusMT Large S+T<sub>ctx</sub> ⑥ mBART-50 S+T<sub>ctx</sub>.

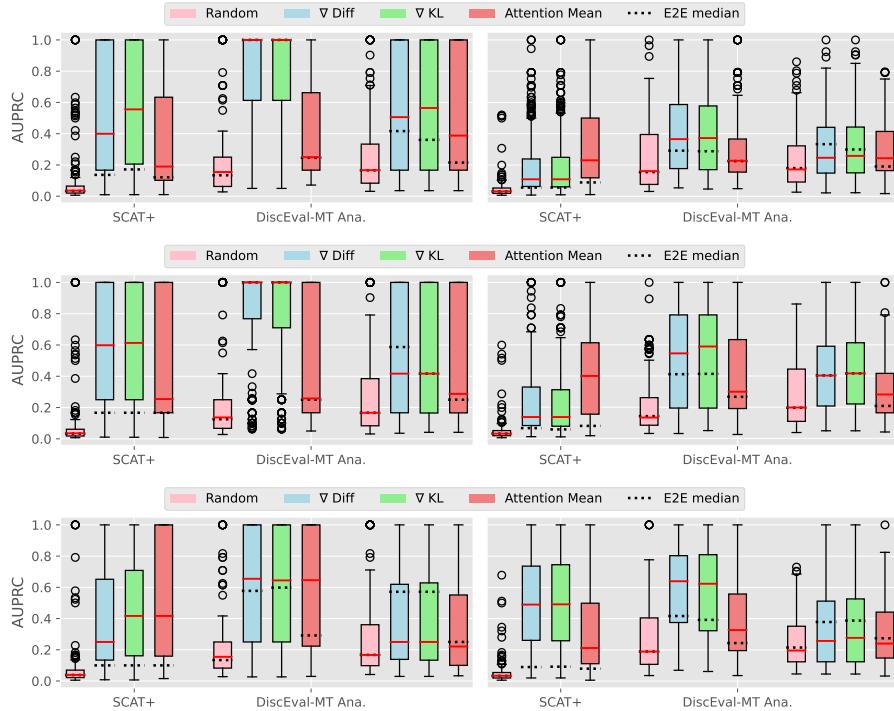
## A. Attributing Context Usage in Multilingual NLP



**Figure A.5 |** Area Under Precision-Recall Curve (AUPRC) of contrastive metrics for context-sensitive target token identification (CTI) on the full datasets (left) or on OK-CS context-sensitive subsets (right). **Top to bottom:** ① OpusMT Small  $S_{ctx}$  ② OpusMT Large  $S_{ctx}$  ③ mBART-50  $S_{ctx}$  ④ OpusMT Small  $S+T_{ctx}$  ⑤ OpusMT Large  $S+T_{ctx}$  ⑥ mBART-50  $S+T_{ctx}$ .



**Figure A.6 |** Macro F1 of CCI methods over full datasets using models trained with only source context (left) or with source+target context (right). Boxes and red median lines show CCI results based on gold context-sensitive tokens. Dotted bars show median CCI scores obtained from context-sensitive tokens identified by KL-Divergence during CTI (E2E settings). **Top to bottom:** ① OpusMT Small  $S_{ctx}$  and  $S+T_{ctx}$  ② OpusMT Large  $S_{ctx}$  and  $S+T_{ctx}$  ③ mBART-50  $S_{ctx}$  and  $S+T_{ctx}$ .



**Figure A.7 |** Area Under Precision-Recall Curve (AUPRC) of CCI methods over full datasets using models trained with only source context (left) or with source+target context (right). Boxes and red median lines show CCI results based on gold context-sensitive tokens. Dotted bars show median CCI scores obtained from context-sensitive tokens identified by KL-Divergence during CTI (E2E settings). **Top to bottom:** ① OpusMT Small S<sub>ctx</sub> and S+T<sub>ctx</sub> ② OpusMT Large S<sub>ctx</sub> and S+T<sub>ctx</sub> ③ mBART-50 S<sub>ctx</sub> and S+T<sub>ctx</sub>.

### A.3 Answer Attribution for Trustworthy Retrieval-Augmented Generation

#### A.3.1 Answer Attribution on the Full XOR-AttriQA

Differently from the concatenation setup in Chapter 5, we also test MIRAGE on the full XOR-AttriQA dataset by constraining CORA generation to match the annotated answer  $y$ . We adopt a procedure similar to Muller et al. (2023) by considering a single document-answer pair  $(\text{doc}_i, y)$  at a time, and using MIRAGE’s CTI step to detect whether  $y$  is sensitive to the context  $\text{doc}_i$ . Results in Table A.3 show that MIRAGE achieves performances in line with other AA methods despite these approaches employing ad-hoc validators trained with as many as 540B parameters.

#### A.3.2 ELI5 Evaluation with Standard Prompt

In the main experiments, we use self-citation prompts by Gao et al. (2023a) for MIRAGE answer attribution to control for the effect of different prompts on model responses, enabling a direct comparison with self-citation. In Table A.5, we provide additional results where a standard prompt without citation instructions is used (“Standard” prompt in Table A.4). We observe the overall citation quality of MIRAGE drops when a standard prompt is used instead of self-citation instructions. We conjecture this might be due to answers that are, in general, less attributable to the provided context due to a lack of explicit instructions to do so. We also observe higher correctness and fluency in the

Method	Extra Requirements	BN	FI	JA	RU	TE	Avg. / Std
mT5 XXL NLI	11B NLI model (250 FT ex.)	81.9	80.9	94.5	87.1	88.7	86.6 / 4.9
	11B NLI model (100k FT ex.)	89.4	88.3	91.5	91.0	92.4	90.5 / 1.5
	11B NLI model (1M FT ex.)	91.1	90.4	93.0	92.9	93.8	92.2 / 1.3
PALM2 <sub>LORA</sub>	540B LLM (250 FT ex.)	91.5	88.3	94.7	93.7	93.7	92.4 / 2.3
PALM2	540B LLM (250 FT ex.)	<b>92.3</b>	<b>92.6</b>	<b>96.4</b>	<b>94.5</b>	<b>94.8</b>	<b>94.1</b> / 1.5
PALM2	540B LLM (4-shot prompting)	91.5	87.4	92.0	90.5	90.6	90.4 / 1.6
PALM2 <sub>CoT</sub>	540B LLM (4-shot prompting)	83.7	78.8	71.7	81.9	84.7	80.2 / 4.7
MIRAGE <sub>CAL</sub> (Ours)	500 AA calibration ex.	<u>82.2</u>	<u>82.5</u>	<u>92.0</u>	<u>87.7</u>	<u>90.2</u>	<u>86.9</u> / 4.0
MIRAGE <sub>EX</sub> (Ours)	-	79.0	74.1	90.8	82.6	86.9	82.7 / 5.8

**Table A.3 |** Agreement % of MIRAGE and entailment-based baselines with human AA on the full XOR-AttriQA using CORA for RAG (annotated answers not matching the LM’s natural generation are force-decoded). **Extra Requirements:** data/models needed for AA in addition to the RAG model itself. **Best overall** and best validator-free scores are highlighted. PALM and mT5 results are taken from Muller et al. (2023).

**SELF-CITATION (Gao et al., 2023a)**

Instruction: Write an accurate, engaging, and concise answer for the given question using only the provided search results (some of which might be irrelevant) and cite them properly. Use an unbiased and journalistic tone. Always cite for any factual claim. When citing several search results, use [1][2][3]. Cite at least one document and at most three documents in each sentence. If multiple documents support the sentence, only cite a minimum sufficient subset of the documents.

**STANDARD**

Instruction: Write an accurate, engaging, and concise answer for the given question using only the provided search results (some of which might be irrelevant). Use an unbiased and journalistic tone.

**Table A.4 |** Self-citation prompt, taken from Gao et al. (2023a), and standard prompt with no citation instruction.

Model	Prompt	Answer Attribution	Filter	Citation↑			Corr.↑	Fluency↑
				Prec.	Rec.	F1		
Zephyr	Self-citation	Self-citation	-	41.4	24.3	30.6	9.9	28.6
		MIRAGE EX	Top 3	38.3	46.2	41.9		
	Standard	MIRAGE EX	Top 5%	<b>44.7</b>	<b>46.5</b>	<b>45.6</b>	<b>11.3</b>	<b>34.3</b>
			Top 3	29.8	34.5	32.0		
LLaMA	Self-citation	MIRAGE EX	Top 5%	34.1	34.2	34.1	11.8	24.8
		Self-citation	-	<b>37.9</b>	19.8	26.0		
	Standard	MIRAGE EX	Top 3	21.8	<b>29.6</b>	25.1	<b>13.0</b>	<b>26.4</b>
			Top 5%	26.2	29.1	<b>27.6</b>		

**Table A.5 |** Citation quality (F1, Recall, Precision), correctness (Corr.), and fluency of self-citation and MIRAGE on ELI5 when using self-citation instruction and standard prompts of Table A.4.

standard prompt setting, suggesting a trade-off between answer and citation quality.

---

INPUT: PROMPT + RETRIEVED DOCS (N=5) + QUERY

---

**Document [2]** [...]

**Document [3]** [...] What will happen if you accidentally set off your security system? The siren will sound and it will be loud, but you should be able to stop the siren by entering your code into your keypad. [...]

**Document [4]** [...]

**Query:** How does a small paper library bar code set off the security alarm when you haven't checked a book out.

---

ANSWER ATTRIBUTION RESULTS

---

**Self-citation:** [...] The response center may contact the individual to confirm if it's a false alarm, and the individual can cancel the alarm by providing their password at the keypad. **[ $\emptyset$ ]** [...]

**MIRAGE:** ... The response **center**<sup>(3)</sup> **may**<sup>(3)</sup> contact the individual to confirm if **it**<sup>(3)</sup>'s a false alarm, and the **individual**<sup>(3)</sup> can **cancel**<sup>(3)</sup> the alarm by **providing**<sup>(3)</sup> their password at the **keypad**<sup>(3)</sup>.**[3]** ...

**NLI (TRUE model):** **[3]** entails the answer sentence.

---

**Table A.6 |** Example described in Chapter 5: MIRAGE attributes the generation to Document [3] when *cancel* is used instead of *prevent*.

### A.3.3 More Examples of Disagreement

Table A.6, Table A.7 and Table A.8 show three cases where MIRAGE answer attributions disagree with self-citation attributions of the same generation.<sup>2</sup> We adopt the Top-5% threshold for CCI Filtering. In Table A.6, the generated answer becomes the consistent description ‘cancel the alarm’ as mentioned in Document [3]. In this case, MIRAGE attributes this sentence to the corresponding Document [3] while NLI maintains its attribution of Document [3] due to lexical overlap, as suggested in Chapter 5.

On several occasions, we observe that MIRAGE attributes all occurrences of lexically similar tokens in the context when the LLM is generating the same word. For example, in Table A.7 the named entity “Science ABC” is mentioned in both Document [1] and [4], and MIRAGE finds both occurrences as salient towards the prediction of the same entity in the output. Similarly, in Table A.8, the generated word ‘Document’ is attributed to the previous mentions of the same word in the context. In both cases, when moving from token-level to sentence-level AA, this dependence would result in

---

<sup>2</sup>Note that we remove citation tags in self-citation generated answers and use MIRAGE to attribute the resulting answers, as introduced in Section 5.5.

wrong AA according to NLI, since the documents are not entailing the answer, but rather making a specific token more likely. These cases reflect the possible discrepancy between AA intended as logical entailment and actual context usage during generation. Future work could explore more elaborate ways to aggregate granular information at sentence level while preserving faithfulness to context usage.

---

**INPUT: PROMPT + RETRIEVED DOCS (N=5) + QUERY**

---

**Document [1]**(Title: Why Do Airlines Sell More Tickets Than The Aircraft's Capacity?  
» Science ABC): [...] They know that some people simply won't show up for their appointments [...] To account for this discrepancy, where the flight would have to fly with some very expensive seats empty, they decide to opt for a more profitable strategy. Assuming that out of the maximum capacity of 200 seats only 185 show up for the flight, this leaves 15 seats available. The airline could try to sell these seats quickly to avoid losses, but this rarely happens. What the airline does instead is try to sell 220 seats in advance.

**Document [2]** [...]

**Document [3]** [...]

**Document [4]**(Title: Why Do Airlines Sell More Tickets Than The Aircraft's Capacity?  
» Science ABC): your own interests, as the airlines have clearly looked after theirs. First of all, if the airline is offering you a seat on a different flight, check if that seat is definitely available. If it is also full and it looks like you might be bumped yet again, you might as well be stranded! If the airline is offering compensation, such as free meals, hotel costs, and transportation between the airport and the hotel, look for the most profitable outcome. The airline might offer you vouchers for future flights, but since airline employees are given guidelines for negotiation, you might

**Document [5]** [...]

---

**Query:** Why are airlines able to sell more seats than are available on a flight?

---

**ANSWER ATTRIBUTION RESULTS**

---

**Self-citation:** [...] According to a report from Science ABC [1], airlines assume that out of the maximum capacity of 200 seats, only 185 will actually board the aircraft. To avoid empty seats and potential losses, airlines opt to sell up to 220 seats in advance [1].  
[...]

**MIRAGE:** [...] According to a report from Science ABC<sup>(1)(4)</sup>, airlines assume that out<sup>(1)</sup> of the maximum capacity<sup>(1)</sup> of 200 seats<sup>(1)</sup>, only 185<sup>(1)</sup> will actually board the aircraft.  
[1][4] [...]

---

**NLI (true model):** Only [1] entails the answer sentence.

---

**Table A.7 |** Example of counterintuitive MIRAGE attribution: the generation is grounded to Document [4] because the named entity Science ABC is made more probable by the occurrence of the same named entity Science ABC. However, Document [4] does not logically entail the answer.

**INPUT: PROMPT + RETRIEVED DOCS (N=5) + QUERY**

**Document [1]**(Title: Why Do Airlines Sell More Tickets Than The Aircraft's Capacity?  
» Science ABC): board the aircraft. They know that some people simply won't show up for their appointments (just like at hospitals, hotels, restaurants etc.). To account for this discrepancy, where the flight would have to fly with some very expensive seats empty, they decide to opt for a more profitable strategy. Assuming that out of the maximum capacity of 200 seats only 185 show up for the flight, this leaves 15 seats available. The airline could try to sell these seats quickly to avoid losses, but this rarely happens. What the airline does instead is try to sell 220 seats in advance.

**Document [2] [...]**

**Document [3] [...]**

**Document [4]**(Title: Why Do Airlines Sell More Tickets Than The Aircraft's Capacity?  
» Science ABC): your own interests, as the airlines have clearly looked after theirs. First of all, if the airline is offering you a seat on a different flight, check if that seat is definitely available. If it is also full and it looks like you might be bumped yet again, you might as well be stranded! If the airline is offering compensation, such as free meals, hotel costs, and transportation between the airport and the hotel, look for the most profitable outcome. The airline might offer you vouchers for future flights, but since airline employees are given guidelines for negotiation, you might

**Document [5] [...]**

**Query:** Why are airlines able to sell more seats than are available on a flight?

**ANSWER ATTRIBUTION RESULTS**

**Self-citation:** [...] According to a source in Document [1], airlines assume that not all passengers will show up for their flights, leaving some seats empty. [...]

**MIRAGE:** [...] According to a source in Document<sup>(4)</sup>, airlines assume that not<sup>(1)</sup> all passengers will show up for their flights, leaving some seats empty. [1][4] [...]

**NLI (TRUE model):** Only [1] entails the answer sentence.

**Table A.8 |** Example of counterintuitive MIRAGE attribution: Document [4] is attributed by MIRAGE due to the repetition of the keyword 'Document'.

## APPENDIX B

---

# Conditioning Generation for Personalized Machine Translation

### B.1 Retrieval and Marking for Attribute-Controlled Translation

#### B.1.1 Prompt Templates

**Formality-Controlled Translation** Here is a sentence: {src} Here is its lang translation written in a attr style: {tgt} The translated sentence conveys a attr style by using words such as ‘w1’, ‘w2’.

**Gender-Controlled Translation** Here is a sentence: {src} Here is its lang translation in which the person is attr: {tgt} In the translation, the attr gender of the person is made explicit by words such as ‘w1’, ‘w2’.

#### B.1.2 Full Per-language Results

- Table B.1: Detailed scores of same-language prompting on CoCoA-MT (preliminary evaluation).<sup>1</sup>
- Table B.2: Full results of same-language prompting on CoCoA-MT (full evaluation).
- Table B.3: Full results of same-language prompting on MT-GENEVAL (full evaluation).
- Table B.4: Full results of cross-lingual prompting on CoCoA-MT.
- Table B.5: Full results of cross-lingual prompting on MT-GENEVAL.

---

<sup>1</sup>We set maximum output length as 50 tokens in the preliminary evaluation, while we use 100 tokens in the main evaluation. Early truncating leads to slightly lower scores in Table B.1 than in Table 6.4.

		BLEU					COMET					Sentential Accuracy					
		0	4	8	16	32	0	4	8	16	32	0	4	8	16	32	
BLOOM 7.1B	base	29	30	31	21	.58	.59	.60	-.09	.76	.84	.88	.73	.56	.76	.84	
	RAMP	31	32	33	25	.60	.61	.62	.15	.84	.89	.91	.84	.44	.86	.90	.92
XGLM 7.5B	base	25	27	28	29	.44	.45	.50	.52	.76	.84	.85	.89	.53	.76	.84	.85
	RAMP	27	28	28	30	.45	.48	.47	.48	.86	.90	.91	.92	.45	.86	.90	.92
GPT- NeoX 20B	base	29	29	29	29	.27	.27	.27	.28	.80	.85	.85	.95	.11	.80	.85	.85
	RAMP	30	30	30	29	.28	.31	.31	.28	.85	.89	.89	.87	.29	.85	.89	.87
BLOOM 175B	base	38	39	39	-	.73	.74	.75	-	.90	.95	.95	-	.48	.74	.76	-
	RAMP	39	40	40	-	.74	.74	.76	-	.95	.97	.97	-	.30	.95	.97	-

**Table B.1** | Detailed scores of same-language prompting on CoCoA-MT (preliminary evaluation). Numbers in the header represent the number of in-context examples used for prompting, including zero-shot prompting (0). Scores are averaged across two available formality values (formal, informal) and languages (ES, FR, HI, PT).

			ES		FR		HI		PT		AVG
			F	I	F	I	F	I	F	I	
XGLM 7.5B	base	BLEU	30.1	33.0	30.7	28.8	18.5	16.9	35.7	35.4	28.6
		COMET	.500	.527	0.348	0.350	0.454	0.425	0.547	0.554	0.463
		L-Acc	.524	.966	.977	0.633	0.976	0.744	0.931	0.928	0.835
		S-Acc	.507	.958	0.953	0.840	0.963	0.748	0.888	0.912	0.846
	+mark	BLEU	31.0	33.2	29.4	27.4	19.2	18.6	35.7	35.5	28.7
		COMET	.498	.541	0.207	0.188	0.439	0.409	0.552	0.552	0.423
		L-Acc	.728	.972	0.985	0.923	0.986	0.860	0.960	0.947	0.920
		S-Acc	.697	.958	0.963	0.917	0.983	0.838	0.927	0.937	0.902
	RAMP	BLEU	32.8	33.5	32.7	31.0	21.0	20.3	34.2	34.4	30.0
		COMET	.480	.511	.314	.302	.502	.491	.488	.522	.451
		L-Acc	.842	.963	.989	.926	.993	.885	.961	.943	.938
		S-Acc	.803	.952	.975	.922	.980	.873	.928	.948	.923
BLOOM 175B	base	BLEU	44.3	45.0	42.9	41.0	27.1	25.8	47.3	45.7	39.9
		COMET	.728	.759	.611	.600	.673	.645	.762	.750	.691
		L-Acc	.795	.960	.987	.890	.978	.885	.987	.954	.930
		S-Acc	.889	.963	.987	.888	.980	.863	.987	.960	.940
	+mark	BLEU	45.8	44.5	43.3	41.8	28.4	27.1	46.4	45.3	40.3
		COMET	.726	.745	.610	.594	.677	.659	.751	.745	.688
		L-Acc	.930	.987	.996	.958	.995	.936	.989	.972	.970
		S-Acc	.942	.985	.992	.957	.992	.925	.990	.977	.970
	RAMP	BLEU	46.4	46.2	43.9	42.9	30.8	29.2	48.8	47.4	41.9
		COMET	.718	.759	.611	.610	.721	.713	.782	.771	.711
		L-Acc	.956	.984	.998	.952	.991	.947	.993	.962	.973
		S-Acc	.957	.982	.995	.945	.993	.935	.990	.967	.970
Adapted MT		BLEU	44.4	43.7	43.4	37.8	19.1	17.0	53.0	49.9	38.5
		COMET	.712	.724	.559	.547	-.191	-.263	.783	.764	.454
		L-Acc	.697	.598	.822	.377	.869	.449	.972	.744	.691
		S-Acc	.700	.600	.810	.400	.680	.600	.950	.800	.693

**Table B.2** | Full results of same-language prompting on CoCoA-MT.

		AR		ES		FR		HI		PT		Avg	
		F	M	F	M	F	M	F	M	F	M	Avg	
XGLM 7.5B	base	BLEU	7.6	7.5	35.5	38.2	27.1	28.6	13.8	16.4	29.2	33.1	23.7
		COMET	-.040	-.012	.694	.738	.509	.555	.304	.332	.661	.713	.445
		L-Acc	.848	.947	.688	.808	.715	.880	.585	.956	.621	.855	.790
		S-Acc	.617	.866	.651	.938	.581	.920	.303	.962	.494	.934	.727
XGLM 7.5B	+mark	BLEU	7.7	7.8	35.4	38.2	27.5	28.7	14.0	16.7	29.1	32.4	23.7
		COMET	-.038	-.020	.704	.735	.508	.556	.300	.317	.663	.714	.444
		L-Acc	.868	.939	.665	.811	.701	.881	.581	.955	.626	.860	.789
		S-Acc	.664	.856	.612	.937	.562	.919	.355	.966	.519	.927	.732
RAMP	base	BLEU	9.2	8.8	37.5	39.4	27.5	29.2	14.8	16.6	31.4	33.3	24.8
		COMET	.037	.043	.723	.759	.528	.571	.325	.337	.681	.723	.473
		L-Acc	.939	.961	.750	.806	.781	.885	.667	.956	.759	.854	.836
		S-Acc	.836	.901	.722	.936	.716	.937	.509	.974	.729	.940	.820
BLOOM 175B	base	BLEU	14.8	16.9	45.6	50.3	38.1	41.7	20.8	24.6	37.6	42.2	33.3
		COMET	.282	.395	.837	.892	.719	.770	.599	.629	.807	.861	.679
		L-Acc	.665	.966	.578	.814	.660	.902	.480	.951	.594	.872	.748
		S-Acc	.411	.934	.515	.965	.581	.961	.212	.973	.525	.960	.704
BLOOM 175B	+mark	BLEU	15.2	17.1	45.8	50.0	37.9	41.3	20.3	23.8	37.6	42.2	33.1
		COMET	.294	.387	.843	.887	.712	.767	.576	.606	.807	.861	.674
		L-Acc	.707	.969	.610	.818	.663	.902	.493	.958	.594	.872	.759
		S-Acc	.482	.936	.568	.973	.588	.962	.284	.974	.525	.960	.725
RAMP	RAMP	BLEU	16.7	17.6	47.9	50.2	39.5	41.8	22.2	25.0	39.3	42.7	34.3
		COMET	.358	.407	.860	.895	.734	.787	.632	.646	.810	.858	.699
		L-Acc	.841	.972	.709	.809	.765	.906	.633	.953	.701	.886	.817
		S-Acc	.721	.940	.707	.964	.732	.971	.518	.973	.683	.972	.818
Adapted MT	MT	BLEU	23.3	24.4	53.2	54.2	44.2	46.4	29.3	32.3	43.4	45.7	35.9
		COMET	.496	.522	.876	.902	.759	.797	.722	.743	.825	.857	.528
		L-Acc	.910	.981	.932	.921	.919	.956	.762	.837	.922	.961	.853
		S-Acc	.940	.970	.910	.960	.950	.960	.280	.750	.930	.990	.863

Table B.3 | Full results of same-language prompting on MT-GENEVAL.

		ES		FR		HI		PT		Avg	
		F	I	F	I	F	I	F	I	Avg	
BLOOM 175B	base	BLEU	40.9	46.3	33.7	32.0	21.8	18.9	33.9	29.0	32.1
		COMET	.785	.823	.611	.615	.409	.436	.772	.705	.644
		L-Acc	.211	.990	.899	.656	.944	.123	.704	.010	.567
		S-Acc	.200	.930	.880	.715	.940	.100	.975	.025	.596
RAMP	RAMP	BLEU	39.4	44.6	35.3	34.7	22.4	18.4	32.2	27.5	31.8
		COMET	.749	.788	.575	.614	.488	.480	.770	.702	.646
		L-Acc	.169	.978	.949	.770	.973	.143	1.00	.015	.625
		S-Acc	.175	.950	.930	.790	.975	.140	.975	.040	.622

Table B.4 | Full results of cross-lingual prompting on CoCoA-MT.

		AR		ES		FR		HI		PT		Avg
		F	M	F	M	F	M	F	M	F	M	
base	BLEU	10.6	11.6	43.3	47.4	34.2	38.2	11.4	15.0	34.4	38.6	28.5
	COMET	.071	.138	.805	.857	.648	.719	-.135	-.003	.766	.822	.469
	L-Acc	.843	.956	.627	.810	.561	.899	.653	.962	.588	.874	.777
	S-Acc	.541	.785	.529	.936	.389	.944	.051	.745	.475	.939	.633
175B	BLEU	10.0	10.5	44.6	47.8	35.7	39.1	13.9	16.6	36.0	39.4	29.4
	COMET	-.044	.020	.818	.860	.686	.739	.139	.212	.779	.816	.502
	L-Acc	.845	.956	.660	.815	.608	.900	.574	.961	.680	.882	.788
	S-Acc	.479	.703	.605	.953	.497	.956	.105	.870	.613	.951	.673

Table B.5 | Full results of cross-lingual prompting on MT-GENEVAL.

### B.1.3 Error Analysis of Cross-Lingual Prompting

Table B.6 shows two examples where RAMP performs significantly worse than the base model in terms of COMET. In the first example, having multiple in-context examples containing “million” led the model to mis-translate “billion” to “million”. In the second example, we observe that the color related in-context examples led the model to produce hallucinated output about clothing colors.

Repeated misleading in-context examples are less observed on MT-GENEVAL and in the same-language setting because (1) CoCoA-MT translates the same set of English sentences to different languages while MT-GENEVAL collects English sentences independently; (2) There are no duplicated source (English) sentences for each language. (Therefore, if RAMP retrieves duplicated English sentences as in Table B.6, their reference translations are guaranteed to be in different languages.)

In-context examples (en)	<ol style="list-style-type: none"> <li>1. Maybe he should. What did you think about that guy findin 3 million dollars worth of old baseball cards in his grandpas attic.</li> <li>2. Yeah that makes sense, did you heard about the \$10 million bunker he has?</li> <li>3. I have. I heard that he started a library in 1895 with 32,000 books in it. All from his personal collection. Can you imagine?</li> <li>4. Yeah that makes sense, did you heard about the \$10 million bunker he has?</li> <li>5. Yeah that makes sense, did you heard about the \$10 million bunker he has?</li> <li>6. Maybe he should. What did you think about that guy findin 3 million dollars worth of old baseball cards in his grandpas attic.</li> <li>7. That is really expensive I agree, did you watch the Lego Batman movie?</li> <li>8. Yeah that makes sense, did you heard about the \$10 million bunker he has?</li> <li>9. That is crazy. Do you like Tom Hanks, he's grossed over 8.5 billion at the box office</li> <li>10. That is really expensive I agree, did you watch the Lego Batman movie?</li> <li>11. That is crazy. Do you like Tom Hanks, he's grossed over 8.5 billion at the box office</li> <li>12. That is crazy. Do you like Tom Hanks, he's grossed over 8.5 billion at the box office</li> <li>13. He doesnt look like he has 56 years! I heard he made 75000000 from Mission Impossible 3</li> <li>14. Really? I guess he made a valuable contribution to science and also to medicine, did you hear of that species of flying snakes</li> </ol>
Input (en) Reference (es) RAMP (es) base (es)	<p>If you got it why not? He is worth over 20 billion dollars after all. Si lo tiene, ¿por qué no? Al fin y al cabo, vale más de 20 000 millones de dólares.</p> <p>Si lo tienes, ¿por qué no? Es millonario después de todo. Si lo tienes, ¿por qué no? Él vale más de 20 mil millones de dólares después de todo.</p>
In-context examples (en)	<ol style="list-style-type: none"> <li>1. thats such a fun thing to have in your code of conduct. are you more of a dog person than cat person?</li> <li>2. For sure lol, it was so nice talking with you, say hi to your cats for me!</li> <li>3. thats such a fun thing to have in your code of conduct. are you more of a dog person than cat person?</li> <li>4. <b>What can't dogs do!</b> I know they aren't color blind like we were taught when young. It was so nice chatting with you, tell yuki hi!</li> <li>5. For sure lol, it was so nice talking with you, say hi to your cats for me!</li> <li>6. Oh yeah. That sucker was mean let me tell you. I think I would have a rabid dog chase me, which by the way dogs are not color blind!</li> <li>7. Thanks you too! And watch out for dogs, they're not actually colorblind but they just don't see as many colors as a normal human can</li> <li>8. LOL That's funny! Google prefers dogs over cats! The wrote in their code of conduct that they are a dog company. It's been fun chatting with you!</li> <li>9. Such gorgeous dogs! Do you &amp; your dogs live in a northern climate?</li> <li>10. LOL That's funny! Google prefers dogs over cats! The wrote in their code of conduct that they are a dog company. It's been fun chatting with you!</li> <li>11. thats such a fun thing to have in your code of conduct. are you more of a dog person than cat person?</li> <li>12. Such gorgeous dogs! Do you &amp; your dogs live in a northern climate?</li> <li>13. haha me too! I heard that they wore clothes that were full of color. Surprised me because I always thought of them as wearing black and white.</li> <li>14. For sure lol, it was so nice talking with you, say hi to your cats for me!</li> </ol>
Input (en) Reference (pt) RAMP (pt) base (pt)	<p>lol I had an African grey that could say everything I said around the house. Like clean your room! as well any loud sounds. "lol Tinha um papagaio-cinzento que conseguia dizer tudo o que ouvia em casa. Tipo ""Limpe o seu quarto!"" e também quisquer sons altos"</p> <p>haha eu também! Eu ouvi dizer que eles usam roupas de todas as cores. Surpreendeu-me porque eu sempre pensei neles usando preto e branco.</p> <p>hahaha eu tinha um papagaio cinza africano que dizia tudo o que eu dizia em casa. Como limpar o quarto! Bem como qualquer som alto.</p>

**Table B.6 |** Examples of CoCoA-MT (formal) where RAMP performs worse than the base model in cross-lingual zero-shot setting. Potentially problematic in-context examples leading to mistranslations or hallucinations are highlighted.

## B.2 Steering Language Models for Machine Translation Personalization

### B.2.1 Experiments Reproducibility

In this section, we provide every parameter we use for the reproducibility of our experiments setups.

#### B.2.1.1 Base Prompt

We use the same prompt template across all methods: ZS (which corresponds to the original model translation), ZS-Exp.<sub>HT</sub>, ZS-Exp.<sub>PT</sub> (detailed in Section B.2.1.4), MS, ActAdd, ReFT, and SAE-based contrastive setups. This prompt, shown in Listing B.3, instructs the model to translate the source sentence while explicitly preventing it from adding any explanations about the translation process. Since all test models are Instruction Tuned, we utilize their native chat templates to preprocess the input accordingly. For multi-shot examples, the *user* and *assistant* turns are repeated for each example, always using the same prompt structure.

#### B.2.1.2 ReFT Training

ReFT training was conducted using the PyReFT toolkit from the original authors<sup>2</sup>. We applied the intervention at the same hook point used by other steering methods – specifically, the layer output corresponding to the residual stream at the selected layer. The training configuration includes a `low_rank_dimension` of 4, `lora_alpha` set to 32, and a `lora_dropout` of 0.05. ReFT was trained on the same 20 prompts used in the MS setup, for a total of 100 epochs.

#### B.2.1.3 SAE Cont.<sub>HT</sub> and SAE Cont.<sub>PT</sub>

We use the NNsight library (Fiotto-Kaufman et al., 2025) to extract and manipulate model activations for all steering experiments. The source code is publicly available in the repository linked in the main body of this paper. For consistency, we use the same set of contrastive examples employed in the MS approach.

#### B.2.1.4 ZS-Exp.<sub>HT</sub> and ZS-Exp.<sub>PT</sub>

For both the ZS-Exp.<sub>HT</sub> and ZS-Exp.<sub>PT</sub> setups, we used GPT-4o (June 2025) to generate explanations detailing the stylistic differences between a base translation and a target

---

<sup>2</sup><https://github.com/stanfordnlp/pyreft>

human translation. The prompt template used for this task is shown in Listing B.1, using the same 20 examples as in the MS, SAE Cont.<sub>HT</sub>, and SAE Cont.<sub>PT</sub> setups.

All outputs were manually inspected to ensure no verbatim excerpts from the provided examples were present, avoiding any risk of data leakage. Example outputs for different novels are shown in Listing B.2.

Finally, these generated guidelines are used to prompt the evaluated models, following the template shown in Listing B.3.

---

**Listing B.1** Prompt template used to get GPT 4o explanation using translation examples.

---

Objective – Identify stylistic choices in translations for  
↪ personalization purposes.

You will be provided with a source text, a standard translation, and a  
↪ target translation by a specific translator whose style we want to  
↪ emulate.

Your task is to analyze the 'Target translation' by comparing it to the  
↪ 'Base translation' and the 'Source text'.

Identify and list the distinctive stylistic patterns, choices, and  
↪ preferences exhibited in the Target translation.

These stylistic cues should help another translator (or an AI) to adapt  
↪ their translations to match the style of the target translator.

Source text: \_\_\_\_\_

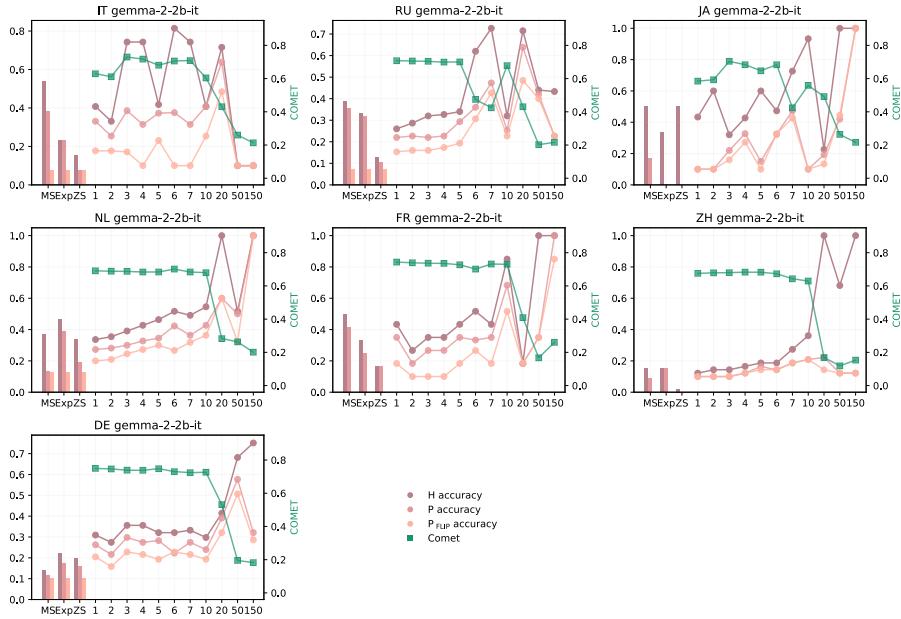
Base translation: \_\_\_\_\_

Target translation: \_\_\_\_\_

[... Repeat Source, Base and Target]

Please extract a concise list of key stylistic cues. Focus on aspects  
↪ such as vocabulary choices, sentence structure, tone and register,  
↪ handling of cultural nuances, punctuation/formatting preferences  
↪ and overall creativity.

Output a short list of stylistic cues as bullet points. Write the list  
↪ as if you were directly giving the guidelines to the translator and  
↪ avoid using specific examples.



**Figure B.1 |** Results for every language on Gemma 2 2B.

## B.2.2 All Models Results

### B.2.2.1 Full Prompting and Steering Results

We present detailed plots of the results for each novel across the three evaluated models in Figure B.1 (Gemma 2 2B), Figure B.2 (Gemma 2 9B), and Figure B.3 (Llama 3.1 8B). These plots display the performance of all evaluated methods, reporting the three submetrics: **H** accuracy (general human-likeness), **P** accuracy (translator-specific accuracy), and **P<sub>flip</sub>** (personalized flip accuracy), alongside the corresponding COMET scores measuring translation quality.

## B.2.3 Dataset and Generation Examples

We present in Table B.7 and Table B.8 a selection of examples from two different languages, showcasing outputs from each of the tested setups. For each example, we also report the corresponding classification label predicted by the classifier and the associated COMET score. Additionally

Source (ZH)	遂又往後看，只著一弓，弓上著香。也有一首歌云：「二十年辨是非，榴花照。三春及初春景，虎兕相逢大。」		
H <sub>1</sub>	Forthwith, therefore, he prosecuted a further perusal of what came next, when he caught sight of a picture of a bow. On this bow hung a citron. There was also this ode: Full twenty years right and wrong to expound will be thy fate! What place pomegranate blossoms come in bloom will face the Palace Gate! The third portion of spring, of the first spring in beauty short will fall! When tiger meets with hare thou wilt return to sleep perennial.	H <sub>1</sub>	-
H <sub>2</sub>	The picture that followed was of a bow with a citron hanging from it, followed by what looked like the words of a song: You shall, when twenty years in life's hard school are done, In pomegranate-time to palace halls ascend. Though three springs never could with your first spring compare, When hare meets tiger your great dream shall end.	H <sub>2</sub>	-
Gemma 2 2B	When he turned back to look, he saw a bow hanging on it, with a plum blossom. There was also a poem that said: "For twenty years, I have been trying to distinguish right from wrong, and the plum blossoms bloom in the palace. In the spring, the tigers and the rhinoceroses meet, and they dream of returning home."	MT	0.49
Gemma 2 9B	Looking back, he saw a painting of a bow with a quiver hanging on it. There was also a line of lyrics that said: For twenty years, I have distinguished right from wrong, and the blossoming of the pomegranate flowers illuminates the imperial court. Three springs have passed, and now it is the beginning of spring again, like a dream where tigers and rhinoceroses meet.	MT	0.56
Llama 3.1 8B	As I looked back, I saw a painting of a bow with a censer on it. There was also a poem that said: Twenty years of distinguishing right from wrong, the lotus flower blooms in the palace. Three springs competing with the scenery of the first spring, the tiger and the rhinoceros meet and return to their dream."	MT	0.53
ZS-Exp-HT	And as he turned his gaze back upon the scene once more, there was displayed an archer's bow strung tautly; upon its surface hung a fragrant plum blossom. A poem also recited thus spake: "For twenty years I have sought truth amidst strife, where blooming durian flowers illuminate the palace walls. In spring's embrace, they vie for beauty, tigers clashing in dreams of return."	MT	0.60
ZS-Exp-PT	And as he turned his gaze back upon the scene once more, there was displayed an archer's bow strung tautly; upon its surface hung a fragrant plum blossom. A poem also recited thus: "For twenty years I have sought truth amidst strife, where blooming durian flowers illuminate the palace walls. In spring's embrace, they vie for beauty, tigers clashing in dreams of return."	MT	0.61
MS	Then, turning around, he noticed a picture depicting a bow hanging on its string adorned with incense sticks. A verse read: "For twenty years I've tried to discern truth from falsehood, and wherever fragrant plum blossoms bloom, the palace gates shine brightly."	H <sub>1</sub>	0.65
ActAdd	When he turned back to look, he saw a bow hanging on it, with a plum blossom. There was also a poem that said: "For twenty years, I have been trying to tell right from wrong, and the plum blossoms bloom in the palace. In spring, the tigers and rhinoceroses meet, and they dream of going back home."	MT	0.53
ReFT	When he turned back to look, he saw a bow hanging there, adorned with a plum blossom. There was also a poem that said: "For twenty years, I have tried to distinguish right from wrong, and plum blossoms bloom within the palace. In spring, tiger and rhinoceros meet, dreaming of their return."	MT	0.58
SAE Cont. <sub>HT</sub> $\alpha = 1$	When he turned back to look, he saw a bow hung upon it, with a plum blossom. There was also a verse which read: "For twenty years, I have struggled to discern truth from falsehood, and the plums blossomed where the imperial court resided. During the three springs, the tiger and rhinoceros met, dreaming of their return."	MT	0.62
$\alpha = 5$	He looked back and there hung a bow adorned with a plum blossom. A verse read: "Twenty years I've struggled for truth, where fragrant plums bloom near the imperial court. Spring arrives, tiger and rhino clash, dreaming of their return."	MT	0.62
$\alpha = 150$ SAE Cont. <sub>PT</sub> $(\alpha = 5)$	The when the veryShellTruth lookWarmTheTheTigeratPoint [...]	MT	0.11
	He looked back and saw a bow adorned with fragrant citron. A verse declared: "Twenty years I've weighed right and wrong; where pomegranate blossoms bloom, they light the palace gates. Spring's full bloom cannot match early spring's grace. When tiger meets rhino, one returns to the great dream."	H <sub>1</sub>	0.65

**Table B.7 |** Example paragraphs from “Dream of the Red Chamber” (ZH→EN) translated with H<sub>1</sub> personalization. Setup outputs are from Gemma 2 2B; only MS and SAE Cont.<sub>PT</sub> successfully flip the classifier prediction.

Source (IT)	Intanto la stalla fu chiusa e Pinocchio rimase solo: e perché erano molte ore che non aveva mangiato, cominciò a sbadigliare dal grande appetito. E, sbadigliando, spalancava una bocca che pareva un forno.		
H <sub>1</sub>	<i>When the stable door was closed, Pinocchio was left alone. And since he hadn't eaten in many hours, he began to yawn from hunger. And when he yawned, his mouth opened as wide as an oven.</i>	H <sub>1</sub>	-
H <sub>2</sub>	<i>The stable was then shut and Pinocchio was left alone. He had not eaten for many hours, and he began to yawn from hunger. And when he yawned he opened a mouth that seemed as wide as an oven.</i>	H <sub>2</sub>	-
Gemma 2 2B	<i>At that moment the stable was closed and Pinocchio was left alone, and because it had been many hours since he had eaten, he began to yawn from a great hunger. And, yawning, he opened a mouth that seemed like a furnace.</i>	MT	0.79
Gemma 2 9B	<i>Meanwhile, the stable was closed and Pinocchio was left alone: and because it had been many hours since he had eaten, he began to yawn from his great hunger. And, yawning, he opened a mouth that looked like an oven.</i>	MT	0.80
Llama 3.1 8B	<i>Meanwhile the stable was closed and Pinocchio was left alone: and because it was many hours since he hadn't eaten, he began to yawn from great hunger. And, yawning, he opened a mouth that seemed like an oven.</i>	MT	0.80
ZS-Exp-HT	<i>Meanwhile, the stable door was shut, leaving Pinocchio alone. His tummy rumbled because it had been hours since his last meal, and he started yawning widely from his hunger. With each yawn, his mouth opened like an oven.</i>	MT	0.78
ZS-Exp-PT	<i>Meanwhile, the stable door was shut, leaving Pinocchio all alone. His tummy rumbled because it had been hours since his last meal, and he started yawning widely from sheer hunger. With each yawn, his mouth gaped open like an oven.</i>	H <sub>2</sub>	0.80
MS	<i>Meanwhile, the stable door was shut and Pinocchio was left alone. Since it had been hours since he last ate, he began yawning from hunger. Each yawn opened his mouth wide enough to look like an oven.</i>	H <sub>2</sub>	0.81
ActAdd	<i>Meanwhile, the stable was closed and Pinocchio remained alone. Because many hours passed without eating, he started yawning from very big hunger. And while yawning, he opened his mouth wide like an oven.</i>	MT	0.75
ReFT	<i>Meanwhile, the stable was closed and Pinocchio stayed alone. Because many hours had passed since he last ate, he began to yawn from big hunger. And, yawning, he opened a mouth that looked like an oven.</i>	MT	0.77
SAE Cont.HT $\alpha = 1$	<i>Meanwhile, the stable door was shut and Pinocchio was left alone. Because it had been many hours since he had eaten, he started to yawn with intense hunger. And, as he yawned, his mouth gaped open like an oven.</i>	MT	0.79
$\alpha = 5$	<i>Meanwhile, the stable door shut, leaving Pinocchio all alone. Because it had been many hours since he'd eaten, he started to yawn with intense hunger. As he yawned, his mouth gaped open, looking like an oven.</i>	H <sub>2</sub>	0.82
$\alpha = 50$	<i>He He "It The ... You It "[They They] THE...I I HE IT [There There THEYRE WE "A A FOR "[...]</i>	MT	0.14
SAE Cont-PT $(\alpha = 5)$	<i>The stable was then shut, leaving Pinocchio all alone. He had not eaten for many hours and began to yawn from great hunger. As he yawned, he opened his mouth wide, like an oven.</i>	H <sub>2</sub>	0.84

**Table B.8 |** Example paragraphs from “Pinocchio” (IT→EN) translated with H<sub>2</sub> personalization. Setup outputs are from Gemma 2 9B; ZS-Exp.PT, MS, SAE Cont.HT and SAE Cont.PT (both  $\alpha = 5$ ) can flip the classifier prediction.

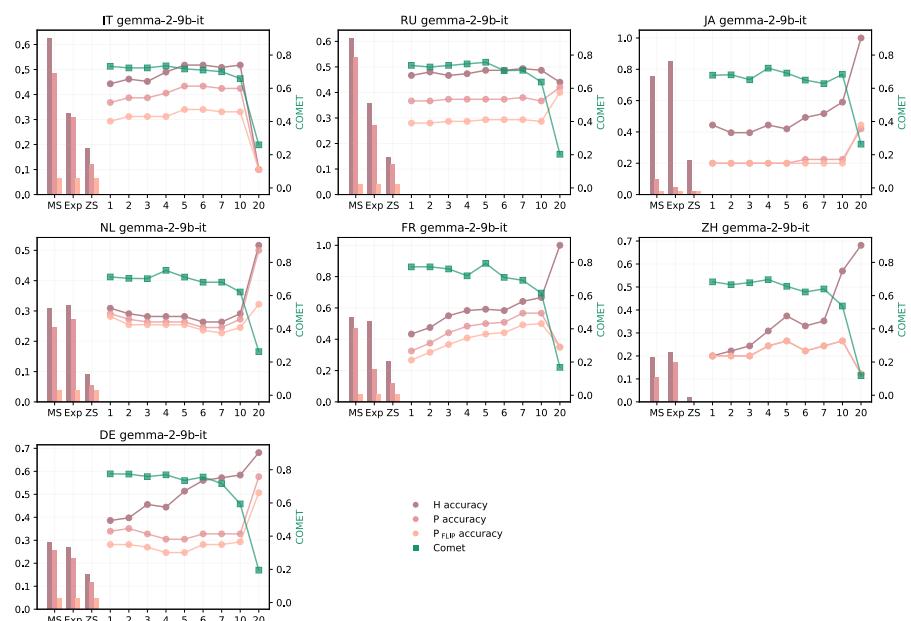


Figure B.2 | Results for every language on Gemma 2 9B.

**Listing B.2** Examples of explanation obtained from GPT 4o when comparing different translations from different novels.

---

[When comparing H1 and MT for Beware of Pity (German)]

- Maintain a tone that is professional, thoughtful, and subtly
  - ↪ persuasive.
- Avoid overly technical jargon unless necessary; explain specialized
  - ↪ terms briefly if used.
- Preserve the author's voice, keeping a balance between academic rigor
  - ↪ and narrative engagement.
- Ensure smooth transitions between sentences and paragraphs to support
  - ↪ coherent argumentation.
- Translate idiomatic expressions in a way that retains their intended
  - ↪ effect, even if the wording differs.

[When comparing H1 and H2 for Beware of Pity (German)]

- Keep the tone warm, welcoming, and direct - avoid overly formal or
  - ↪ distant language.
- Use active voice wherever possible to maintain energy and engagement.
- Avoid idioms or expressions that may not translate culturally; aim for
  - ↪ universal accessibility.
- Maintain consistent tone and register throughout, adapting to the
  - ↪ intended audience's familiarity with the subject.
- Respect the rhythm and structure of the original, but feel free to
  - ↪ adjust sentence length for readability.

[When comparing H1 and H2 for Crime and Punishment (Russian)]

- Prefer dynamic over formal vocabulary: Opt for vivid or emotionally
  - ↪ charged words when available.
- Add tonal nuance and emotional shading: Enrich dialogues and narration
  - ↪ with subtle shifts in tone, especially sarcasm, understatement, or
  - ↪ irony, to match character voice or mood.
- Use contractions and familiar phrasing: Employ contractions and
  - ↪ relaxed expressions to preserve spoken character.
- Expand or rephrase for clarity and voice: Don't hesitate to slightly
  - ↪ reword or elaborate if it strengthens tone, clarifies intent, or
  - ↪ enhances character differentiation.
- Favor rhythmic, flowing sentence structure: Break long, formal
  - ↪ sentences into multiple shorter clauses or use punctuation (dashes,
  - ↪ ellipses) for dramatic or emotional effect.

[...]

---

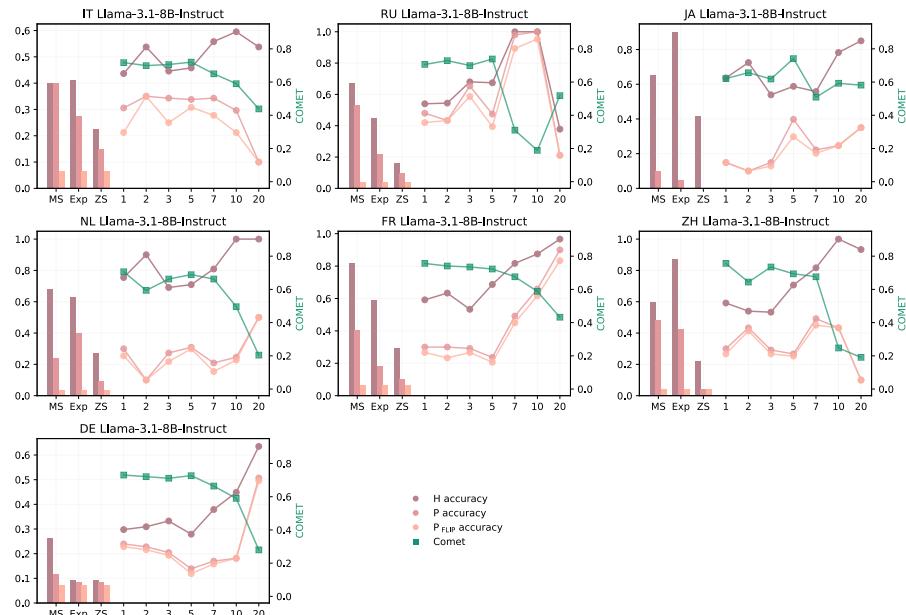
**Listing B.3** Zero shot template template when prompting language models with different setups

Translate the following sentence between the angular parentheses into  
 $\hookrightarrow$  English.

```
[if setup == ZS-Exp]
[Follow the following guidelines when translating: <explanations here>]
```

The original sentence is: \_\_\_\_\_.

Remember to write only the translation, without any additional text or  
 $\hookrightarrow$  explanation.



**Figure B.3 |** Results for every language on Llama 3.1 8B.



# APPENDIX C

---

## Interpretability in Human Translation Workflows

### C.1 Machine Translation Post-editing for Typologically Diverse Languages

#### C.1.1 Subject Information

During the setup of our experiment, one translator refused to carry out the main task after the warmup phase, and another was substituted by our choice. Both translators were working in the English–Italian direction and were found to make heavy usage of copy-pasting during the warmup stage, suggesting an incorrect utilization of the platform in light of our guidelines. Both translators, which we identified as  $T_2$  and  $T_3$  for Italian, were replaced by  $T_5$  and  $T_4$  respectively. Table C.1 reflects the final translation selection for all languages, with the information collected by means of the pre-task questionnaire.

#### C.1.2 Translation Guidelines

An extract of the translation guidelines provided to the translators follows. The full guidelines are provided in the additional materials.

Fill in the pre-task questionnaire before starting the project. In this experiment, your goal is to complete the translation of multiple files in one of two possible translation settings. Please, complete the tasks on your own, even if you know another translator that might be working on this project. The translation setting alternates between texts, with each text requiring a single translation in the assigned setting. The two translation settings are:

		Gender	Age	Degree	Position	En Level	YoE	PE YoE	% PE
<b>Arabic</b>	T <sub>1</sub>	M	35-44	BA	Freelancer	C2	> 15	2-5	20-40%
	T <sub>2</sub>	M	25-34	BA	Employed	C2	5-10	2-5	60-80%
	T <sub>3</sub>	M	25-34	MA	Freelancer	C1	5-10	< 2	20-40%
<b>Dutch</b>	T <sub>1</sub>	M	25-34	MA	Freelancer	C2	5-10	5-10	60-80%
	T <sub>2</sub>	F	35-44	MA	Freelancer	C1	10-15	5-10	40-60%
	T <sub>3</sub>	F	25-34	MA	Freelancer	C2	2-5	2-5	20-40%
<b>Italian</b>	T <sub>1</sub>	F	25-34	MA	Employed	C1	5-10	5-10	20-40%
	T <sub>5</sub>	F	25-34	MA	Freelancer	C1	2-5	2-5	40-60%
	T <sub>4</sub>	F	35-44	BA	Freelancer	C2	10-15	5-10	>80%
<b>Turkish</b>	T <sub>1</sub>	F	25-34	BA	Freelancer	C2	5-10	2-5	<20%
	T <sub>2</sub>	F	25-34	BA	Freelancer	C1	5-10	5-10	<20%
	T <sub>3</sub>	M	25-34	High sch.	Freelancer	C2	10-15	< 2	<20%
<b>Ukrainian</b>	T <sub>1</sub>	F	35-44	MA	Employed	C1	5-10	5-10	20-40%
	T <sub>2</sub>	M	35-44	MA	Employed	C1	10-15	10-15	20-40%
	T <sub>3</sub>	M	35-44	High sch.	Employed	B2	2-5	2-5	20-40%
<b>Vietnamese</b>	T <sub>1</sub>	F	25-34	MA	Employed	C2	10-15	5-10	40-60%
	T <sub>2</sub>	F	25-34	BA	Freelancer	C1	5-10	< 2	20-40%
	T <sub>3</sub>	F	25-34	MA	Employed	C1	2-5	< 2	<20%

**Table C.1 |** Subjects information for DivEMT. The last three columns represent respectively the number of years of professional experience as a translator (YoE), the number of years of experience with MT post-editing (PE YoE) and the % of work assignments requiring post-editing in the last 12 months (% PE) for each subject.

- Translation from scratch.** Only the source sentence is provided, you are to write the translation from scratch.
- Post-editing.** The source sentence is provided alongside a translation produced by an MT system. You are to post-edit this MT output. Post-edit the text so you are satisfied with the final translation (the required quality is publishable quality). If the MT output is too time-consuming to fix, you can delete it and start from scratch. However, please do not systematically delete the provided MT output to give your own translation.

**Important:** All editing MUST happen in the provided PET interface: that is, working in other editors and copy-pasting the text back to PET is NOT ALLOWED, because it invalidates the experiment. This is easy to spot in the log data, so please avoid doing this. Complete the translation of all files

sequentially, i.e. in the order presented in the tool. DO NOT SKIP files at your own convenience. Make sure that ALL files are translated when you deliver the tasks.

The aim is to produce publishable professional quality translations for both translation settings. Thus, please translate to your best abilities. You can return to the files and self-review as many times as you think it is necessary. Important: The time invested to translate is recorded while the active unit (sentence) is in editing mode (yellow background). Therefore:

- Only start to translate when you are in editing mode (yellow background). In other words, do not start thinking how you will translate a sentence when the active unit is not yet in editing mode (green or red background).
- Do not leave a unit in editing mode (yellow background) while you do something else. If you need to do something unrelated in the middle of a translation then go out of editing mode and come back to editing mode when you are ready to resume translating.
- First you will be translating a warmup task, and then the main task. When you are translating each file, you can consult the source text by looking up the url in the Excel files that we have sent for reference.

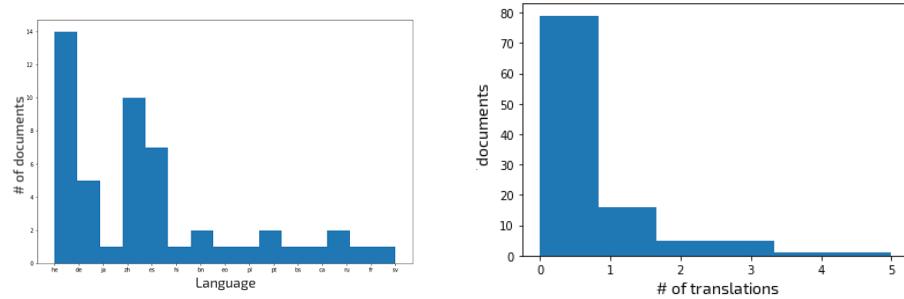
In order to find the correct terminology for the translation you can consult any source in the Internet. Important: However, it is **NOT ALLOWED** to use any MT engine to find terms or alternatives to translations (such as Google Translate, DeepL, MS Translator or any MT engine available in your language). Using MT engines invalidates the experiment, and will be detected in the log data. Please fill-in the post-task questionnaire ONLY ONCE after completing all the translation tasks (both warmup and main tasks).

### C.1.3 Details on Document Selection and Preprocessing

**Document selection** Table C.2 present the distribution of selected documents from the Flores-101 devtest split based on their domain and the number of sentences that compose them. The first goal in the selection process was to preserve a rough balance between the three categories while including mostly 4 and 5-sentence docs which are faster to edit in PET (no need to frequently close and reopen an editing window). Another objective of the selection was to minimize the chance of translators finding

Type	WN	WV	WB	# Sent.	# Words
3S	11	13	11	105	2168
4S	14	8	13	140	3214
5S	12	13	12	185	3826
Tot.	37	34	36	450	9626

**Table C.2 |** Distribution of the selected DivEMT documents across sizes and Wikipedia categories. A Type value of  $N$ S stands for documents composed by  $N$  contiguous sentences, WN, WV and WB stand respectively for WikiNews, WikiVoyage and Wikibooks



**Figure C.1 | Left:** Distribution for the availability of documents selected for DivEMT in languages other than English. **Right:** Quantity of selected documents per number of available translations of Wikipedia.

the translated version of the Wikipedia article from which documents were taken and copied from there, despite our guidelines. We thus scrape the articles from Wikipedia and assess the number of available translations. Among the selected documents, only a small subset has translations in other languages (see Figure C.1 top, an article can have multiple languages), mainly in Hebrew (14), Chinese (10), Spanish (7) and German (5) respectively. Considering the total number of translations for every article (Figure C.1 bottom), we see that roughly 75% of them (79 docs) have no translations. We consider this satisfactory as proof there should not be a large amount of possible copying involved, and we follow up on this evaluation by also ensuring that no repeated copy-paste patterns are present in keylogs after the warmup stage.

**Filtering of Outliers** For our analysis of Section 8.4, we only use sentences with an editing time lower than 45 minutes, which was selected heuristically as a reasonably high threshold to allow for extensive searching and thinking. In the following, we present the identifiers of the sentences that were filtered out during this process. E.g. 54.1 means

the first sentence of document 54, having `item_id` equal to `flores101-main-541` in the dataset. Note that the sentences were outliers only for 2/6 languages and were all different, indicating no systematic issues in the sample: ARA: 54.1, 100.3, VIE: 3.1, 3.2, 24.3, 28.4, 33.1, 33.2, 40.3, 41.2, 50.3, 100.1, 102.1, 106.1, 107.2, 107.4. The 17 sentences were removed for all modalities and languages in the analysis of Section 8.4 to preserve the validity of our comparison, representing a loss of roughly 4% of the total available data, a tolerable amount for our analysis.

**Fields Description** Table C.3 presents the set of fields that were collected for every entry of the DivEMT dataset. The fields related to keystrokes, times, pauses, annotations and visit order were extracted from the event log of PET .per files, while edits information and other MT quality metrics were computed in a second moment with the help of widely-used libraries.

#### C.1.4 Other Measurements

**Automatic Evaluation of NMT Systems** The selection of systems used in this study was driven by a broader evaluation procedure covering more models, metrics and target languages. Table C.4 presents the overall results of our evaluation. We use HuggingFace’s `transformers` library (Wolf et al., 2020) for all neural models, using the default decoding settings without further fine-tuning. All metrics were computed using the default settings of SacreBLEU (Post, 2018) and COMET (Rei et al., 2020).

**Inter-subject Variability in Translation Times** Although the variability across different subjects working on the same language directions is not the main concern of our investigation, we produce Figure C.2 (an expanded version of Figure 8.2) to visualize the inter-subject variability for translation times. We observe that the variability across different translators is more pronounced when translating from scratch and that the overall trend of speed improvements associated with PE is mostly preserved (with few exceptions related to the PE<sub>2</sub> modality).

Field name	Description
<code>unit_id, flores_id, subject_id, task_type</code>	Identifiers for the item, respective FLORES-101 sentence, translator and translation mode.
<code>src_text</code>	The original source sentence extracted from Wikinews, wikibooks or wikivoyage.
<code>mt_text</code>	MT output sentence before post-editing, present only if <code>task_type</code> is 'pe'.
<code>tgt_text</code>	Final sentence produced by the translator (either from scratch or post-editing <code>mt_text</code> )
<code>aligned_edit</code>	Aligned visual representation of the machine translation and its post-edit with edit operations
<code>edit_time</code>	Total editing time for the translation in seconds.
<code>k_letter, k_digit, k_white, k_symbol, k_nav</code>	Number of keystrokes for various key types (letters, digits, keystrokes, whitespaces, punctuation, navigation keys) during the translation.
<code>k_eraser, k_copy, k_paste, k_cut, k_do</code>	Number of keystrokes for erase (backspace, cancel), copy, paste, cut and Enter actions during the translation.
<code>k_total</code>	Total number of all keystroke categories during the translation.
<code>n_pause_geq_N, len_pause_geq_N</code>	Number and length of pauses longer than 300ms and 1000ms during the translation.
<code>num_annotations</code>	Number of times the translator focused the target sentence textbox during the session.
<code>n_insert, n_delete, n_substitute, n_shift, tot_shifted_words, tot_edits, hter</code>	Granular editing metrics and overall HTER computed using the Tercom library.
<code>cer</code>	Character-level HTER score computed between the MT and post-edited outputs.
<code>bleu, chrf</code>	Sentence-level BLEU and ChrF scores between MT and post-edited fields computed using the SacreBLEU library with default parameters.
<code>time_per_char, key_per_char, words_per_hour/minute</code>	Edit time per source character, expressed in seconds. Proportion of keys per character needed to perform the translation. Amount of source words translated or post-edited per hour/minute
<code>subject_visit_order</code>	Id denoting the order in which the translator accessed documents in the interface.

**Table C.3 |** Description of the main fields associated to every DivEMT data entry. An entry correspond to a translation in a specific modality (HT, PE<sub>1</sub> or PE<sub>2</sub>) for one of the six target languages

	System	BLEU	chrF2	TER	chrF2++	COMET
<b>Arabic</b>	M2M100	19.2	50.9	69.2	47	0.417
	MarianNMT	<u>22.7</u>	<u>54.2</u>	<u>64.7</u>	<u>50.4</u>	0.483
	mBART-50	<u>17</u>	48.5	69.1	44.8	0.452
	GTrans	<b>34.1</b>	<b>65.6</b>	<b>52.8</b>	<b>61.9</b>	<b>0.737</b>
<b>Dutch</b>	M2M100	21.3	52.9	66.1	49.8	0.405
	MarianNMT	<u>25</u>	<u>56.9</u>	<u>62.5</u>	<u>53.8</u>	0.543
	mBART-50	22.6	53.9	63.7	50.9	0.532
	DeepL	28.7	59.5	59.5	56.6	<b>0.67</b>
	GTrans	<b>29.1</b>	<b>60</b>	<b>58.5</b>	<b>57.1</b>	0.667
<b>Indonesian</b>	M2M100	35.9	63.1	47.3	60.8	0.614
	MarianNMT	<u>38.5</u>	<u>65.6</u>	<u>46.5</u>	<u>63.3</u>	0.671
	mBART-50	<u>35.9</u>	<u>63.3</u>	<u>47.7</u>	<u>61.1</u>	0.706
	GTrans	<b>51.5</b>	<b>73.6</b>	<b>34.5</b>	<b>71.9</b>	<b>0.894</b>
<b>Italian</b>	M2M100	23.6	53.9	63.2	51	0.51
	MarianNMT	<u>27.5</u>	<u>57.6</u>	<u>58.9</u>	<u>54.8</u>	0.642
	mBART-50	24.4	54.7	61.2	51.8	0.648
	DeepL	<u>33</u>	61	54	58.5	<b>0.795</b>
	GTrans	32.8	<b>61.4</b>	<b>53.6</b>	<b>58.8</b>	0.781
<b>Japanese</b>	M2M100	24.5	32.2	123.3	26	0.389
	mBART	<u>27.1</u>	<u>35.4</u>	<u>123</u>	<u>28.3</u>	0.538
	DeepL	<u>41.3</u>	<b>46.8</b>	108	<u>37</u>	<b>0.75</b>
	GTrans	38.4	44.7	<b>101.5</b>	33.9	0.683
<b>Polish</b>	M2M100	16.1	46.5	74.2	43.1	0.486
	MarianNMT	<u>19.3</u>	<u>49.9</u>	<u>70.5</u>	<u>46.6</u>	0.648
	mBART-50	17.4	48.2	72.4	44.9	0.603
	DeepL	24	54.3	66.4	51.1	<b>0.832</b>
	GTrans	<b>24.4</b>	<b>54.6</b>	<b>64.6</b>	<b>51.4</b>	0.804
<b>Russian</b>	M2M100	22.5	51.1	65.6	48.1	0.427
	MarianNMT	<u>25.4</u>	<u>53.5</u>	64.3	<u>50.7</u>	0.537
	mBART	24.8	52.6	<u>63.7</u>	49.7	<u>0.541</u>
	DeepL	<u>35.9</u>	<b>61.8</b>	<u>53.3</u>	<b>59.3</b>	<b>0.79</b>
	GTrans	33	60.5	55.2	57.7	0.731
<b>Turkish</b>	M2M100	20.3	53.9	65.2	50.1	0.686
	MarianNMT	<u>26.3</u>	<u>59.8</u>	<u>58.8</u>	<u>55.8</u>	<u>0.881</u>
	mBART-50	18.8	52.7	<u>67.5</u>	48.7	0.755
	GTrans	<u>35</u>	<b>65.5</b>	<b>50.4</b>	<b>62.2</b>	<b>1</b>
<b>Ukrainian</b>	M2M100	21.9	51.4	65.8	48.3	0.463
	MarianNMT	<u>20</u>	48.8	69.2	45.7	0.427
	mBART-50	<u>21.9</u>	<u>50.7</u>	<u>67.9</u>	<u>47.7</u>	<u>0.587</u>
	GTrans	<u>31.1</u>	<u>59.8</u>	<u>55.9</u>	<u>56.8</u>	<u>0.758</u>
<b>Vietnamese</b>	M2M100	33.3	52.3	52.4	52.1	0.43
	MarianNMT	26.7	45.7	60.2	45.6	0.117
	mBART-50	<u>34.7</u>	<u>54</u>	<u>50.7</u>	<u>53.8</u>	0.608
	GTrans	<b>45.1</b>	<b>61.9</b>	<b>41.8</b>	<b>61.9</b>	<b>0.724</b>

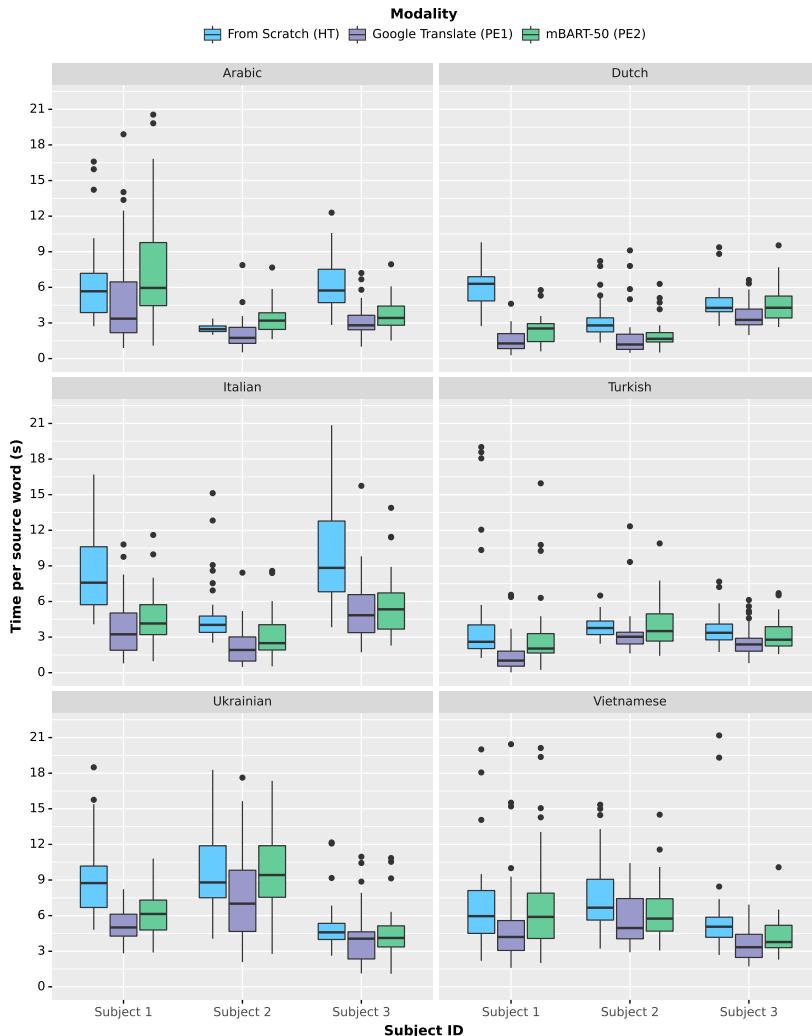
**Table C.4 |** Automatic MT quality of all evaluated NMT systems on all tested languages in the English-to-XX setting, using the FLORES-101 full devtest for evaluation. Besides mBART-50 and Google Translate (GTrans), we also evaluate a set of bilingual Transformer-based NMT models trained with MarianNMT (Tiedemann and Thottingal, 2020), the DeepL industrial MT system and the multilingual M2M-100 418M model (Fan et al., 2021). Best overall and open-source only performances are highlighted.

	ENGLISH	Inland waterways can be a good theme to base a holiday around.
ARABIC	HT	يمكن أن تكون المارات المائية الداخلية خياراً جيداً لخطط عطلة حولها.
	PE <sub>1</sub>	يمكن أن تكون المارات المائية الداخلية موضوعاً جيداً لإقامة عطلة حولها. MT: يمكن أن تكون المارات المائية الداخلية مثلاً [جيّداً] لإقامة عطلة حولها.
	PE <sub>2</sub>	يمكن أن تكون السكك الحديدية الداخلية موضوعاً جيداً لإقامة عطلة حولها. قد تكون المارات المائية الداخلية مكاناً جيداً لقضاء عطلة حولها. PE: قد تكون المارات المائية الداخلية حولها.
DUTCH	HT	Binnenlandse waterwegen kunnen een goed thema zijn voor een vakantie.
	PE <sub>1</sub>	MT: De binnenwateren kunnen een goed thema zijn om een vakantie omheen te baseren. PE: Binnenwateren kunnen een goede vakantiebestemming zijn.
	PE <sub>2</sub>	MT: Binnenwaterwegen kunnen een goed thema zijn om een vakantie rond te zetten. PE: Binnenwaterwegen kunnen een goed thema zijn om een vakantie rond te organiseren.
ITALIAN	HT	I corsi d'acqua dell'entroterra possono essere un ottimo punto di partenza da cui organizzare una vacanza.
	PE <sub>1</sub>	MT: Trasporto fluviale può essere un buon tema per basare una vacanza in giro. PE: I canali di navigazione interna possono essere un ottimo motivo per cui intraprendere una vacanza.
	PE <sub>2</sub>	MT: I corsi d'acqua interni possono essere un buon tema per fondare una vacanza. PE: I corsi d'acqua interni possono essere un buon tema su cui basare una vacanza.
TURKISH	HT	İç bölgelerdeki su yolları, tatil planı için iyi bir tema olabilir.
	PE <sub>1</sub>	MT: İç su yolları, bir tatili temel almak için iyi bir tema olabilir. PE: İç su yolları, bir tatil planı yapmak için iyi bir tema olabilir.
	PE <sub>2</sub>	MT: İç su yolları, tatil için uygun bir tema olabilir. PE: İç su yolları, tatil için uygun bir tema olabilir.
UKRAINIAN	HT	Можна спланувати вихідні, взявшися за основу подорож внутрішніми водними шляхами.
	PE <sub>1</sub>	MT: Внутрішні водні шляхи можуть стати гарною темою для відпочинку навколо. PE: Внутрішні водні шляхи можуть стати гарною темою для проведення вихідних.
	PE <sub>2</sub>	MT: Водні шляхи можуть бути хорошим об'єктом для базування відпочинку навколо. PE: Місцевість навколо внутрішніх водних шляхів може бути гарним вибором для організації відпочинку.
VIETNAMESE	HT	Du lịch trên sông có thể là một lựa chọn phù hợp cho kỳ nghỉ.
	PE <sub>1</sub>	MT: Dường thủy nội địa có thể là một chủ đề hay để tạo cơ sở cho một kỳ nghỉ xung quanh. PE: Đường thủy nội địa có thể là một ý tưởng hay để lập kế hoạch cho kỳ nghỉ.
	PE <sub>2</sub>	MT: Các tuyến nước nội địa có thể là một chủ đề tốt để xây dựng một kì nghỉ. PE: Du lịch bằng đường thủy nội địa là một ý tưởng nghỉ dưỡng không tồi.

**Table C.5 |** An example sentence (81.1) from the DivEMT corpus, with the English source and all output modalities for all target languages, including intermediate machine translations (MT) and subsequent post-editings (PE). Colors denote **insertions**, **deletions**, **substitutions** and **shifts** computed with Tercom (Snover et al., 2006).

	ENGLISH	The Internet combines elements of both mass and interpersonal communication.
	ARABIC	يجمع الانترنت بين عناصر وسائل الاتصال العامة والشخصية على حد سواء.
	HT	يجمع الانترنت بين عناصر وسائل الاتصال العامة والشخصية على حد سواء.
	PE1	<b>يجمع</b> الانترنت بين عناصر الاتصال الجماعي والشخصي: PE: <b>يجمع</b> الانترنت بين عناصر الاتصال الجماعي والشخصي
	PE2	إنترنت <b>يجمع</b> عناصر التواصل الجماعي والتواصل الشخصي: MT: إنترنت <b>يجمع</b> عناصر التواصل الجماعي والتواصل الشخصي PE: <b>يجمع</b> شبكة الانترنت عناصر التواصل الجماعي والتواصل الشخصي
	DUTCH	Het internet combineert elementen van zowel massa- en intermenselijke communicatie.
	HT	Het internet combineert elementen van zowel massa- als interpersoonlijke communicatie.
	PE1	<b>MT:</b> Het internet combineert elementen van zowel massa- als interpersoonlijke communicatie. PE: Het internet combineert elementen van zowel massa- als interpersoonlijke communicatie.
	PE2	<b>MT:</b> Het internet combineert elementen van massa- en interpersoonlijke communicatie. PE: Binnenwaterwegen Het internet combineert elementen van massa- en interpersoonlijke communicatie.
	ITALIAN	Internet combina elementi di comunicazione sia di massa sia interpersonale.
	HT	Internet combina elementi di comunicazione sia di massa sia interpersonale.
	PE1	<b>MT:</b> Internet combina elementi di comunicazione di massa e <b>interpersonali</b> . PE: Internet combina elementi di comunicazione di massa e <b>interpersonale</b> .
	PE2	<b>MT: MT:</b> Internet combina elementi di comunicazione di massa e interpersonale. PE: Internet combina elementi di comunicazione di massa e interpersonale.
	TURKISH	İnternet hem kitlesel hem de bireysel iletişim öğelerini birleştiriyor.
	HT	İnternet hem kitlesel hem de bireysel iletişim öğelerini birleştiriyor.
	PE1	<b>MT:</b> İnternet, hem <b>kitle</b> hem de kişiler arası iletişimini unsurlarını birleştirir. PE: İnternet, hem <b>kitleler</b> hem de kişiler arası iletişimini unsurlarını birleştirir.
	PE2	<b>MT:</b> İnternet hem kitlesel hem de kişisel iletişim unsurlarını birleştiriyor. PE: İnternet hem kitlesel hem de kişisel iletişim unsurlarını birleştiriyor.
	UKRAINIAN	В інтернеті поєднуються елементи групового спілкування та особистого спілкування.
	HT	В інтернеті поєднуються елементи групового спілкування та особистого спілкування.
	PE1	<b>MT:</b> Інтернет поєднує в собі елементи як масового, так і міжособистісного спілкування. PE: Інтернет поєднує в собі елементи як масового, так і міжособистісного спілкування.
	PE2	<b>MT:</b> Інтернет <b>об 'єднує як масову</b> , так і <b>міжлюдську комунікацію</b> . PE: Інтернет <b>поєднує в</b> собі елементи як <b>групової</b> , так і <b>особистої комунікації</b> .
	VietNAMESE	Internet là nơi tổng hợp các yếu tố của cả phương tiện truyền thông đại chúng và giao tiếp liên cá nhân.
	HT	Internet là nơi tổng hợp các yếu tố của cả phương tiện truyền thông đại chúng và giao tiếp liên cá nhân.
	PE1	<b>MT:</b> Internet kết hợp các yếu tố của cả giao tiếp đại chúng và giao tiếp giữa các cá nhân. PE: Internet kết hợp các yếu tố của cả truyền thông đại chúng và giao tiếp giữa các cá nhân.
	PE2	<b>MT:</b> Internet kết hợp những yếu tố của sự giao tiếp quản chúng và giao tiếp giữa người với người. PE: Internet kết hợp những yếu tố của cả việc giao tiếp đại chúng và giao tiếp cá nhân.

**Table C.6 |** An example sentence (29.2) from the DivEMT corpus, with the English source and all output modalities for all target languages, including intermediate machine translations (MT) and subsequent post-editings (PE). Colors denote **insertions**, **deletions**, **substitutions** and **shifts** computed with Tercom (Snover et al., 2006).



**Figure C.2 |** Time per processed source word across languages, subjects and translation modalities, measured in seconds. Each point represents a document containing 3–5 sentences translated by a subject in one of the languages, with higher scores representing slower editing.

Subject	Coefficient
ara_t1	0.281
ara_t2	-0.384
ara_t3	-0.103
nld_t1	0.001
nld_t2	-0.459
nld_t3	0.458
ita_t1	0.086
ita_t4	0.350
ita_t5	-0.436
tur_t1	-0.381
tur_t2	0.272
tur_t3	0.109
ukr_t1	0.077
ukr_t2	0.314
ukr_t3	-0.391
vie_t1	0.012
vie_t2	0.176
vie_t3	-0.188

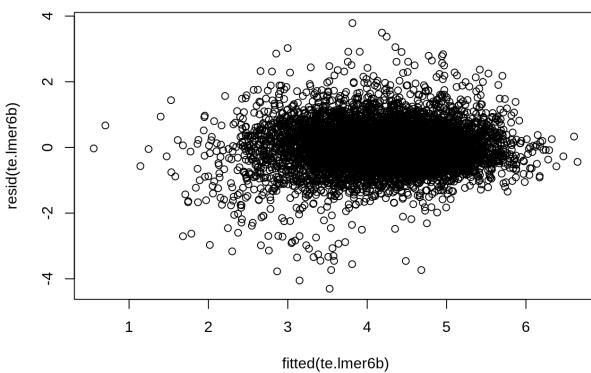
**Table C.7 |**Coefficients of the random intercept related to the `subject_id` variable, representing the identity of the translator performing the translation.

### C.1.5 Data Filtering and Feature Significance

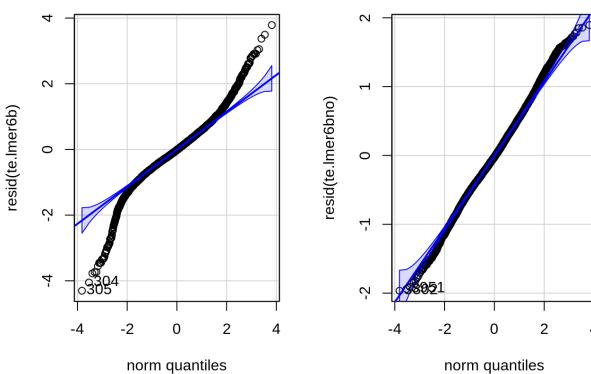
We log-transform the dependent variable, edit time in seconds, given its long right tail. The models are built by adding one element at a time, and checking whether such addition leads to a significantly better model with AIC (i.e. if the score gets reduced by at least 2). Our random effects structure includes random intercepts for different segments (nested with documents) and translators, as well as a random slope for modality over individual segments. We start with an initial model that just includes the two random intercepts (by-translator and by-segment) and proceed by (i) finding significance for nested document/segment random effect; (ii) adding fixed predictors one by one; (iii) adding interactions between fixed predictors; and (iv) adding the random slopes.<sup>a</sup> From this sequential procedure, we obtain the resulting model. When checking the homoscedasticity and normality of residuals assumptions (Figure C.3 and Figure C.4), we find the latter is not fulfilled. Consequently, we remove data points for which observations deviate by more than 2.5 standard deviations from the predicted value by the model (2.4% of the data) and refit the best model on this subset, in order to find out whether any of the effects were due to these outliers. The resulting trends do not change significantly in this final model, in which residuals are normally distributed. As a final sanity check, in Table C.7 we measure the effect of subject identity on edit times and find no systematic patterns across languages.

---

<sup>a</sup>The document processing order was originally included to identify possible longitudinal effects but was removed due to a lack of significant improvements.



**Figure C.3 |** Residuals of the final LMER model, used to verify the heteroscedasticity assumption.



**Figure C.4 |** Quantile-quantile plot before and after the removal of outliers when fitting the LMER model, used to verify the normality assumption.

## C.2 Word-level Quality Estimation for Machine Translation Post-editing

### C.2.1 Filtering Details for QE4PE Data

1. *Documents should contain between 4 and 10 segments, each containing 10–100 words (959 docs).* This ensures that all documents are roughly uniform in terms of size and complexity to maintain a steady editing flow Section 9.2.5.
2. *The average segment-level QE score predicted by XCOMET-XXL is between 0.3 and 0.95, with no segment below 0.3 (429 docs).* This forces segments to have a decent but still imperfect quality, excluding fully wrong translations.
3. *At least 3 and at most 20 errors spans per document, with no more than 30% of words in the document being highlighted (351 docs).* This avoids overwhelming the editor with excessive highlighting, while still ensuring error presence.

The same heuristics were applied to both translation directions, selecting only documents matching our criteria in both cases.

### C.2.2 Additional Details and Statistics

ID	Job	Eng. Lvl	Trans. YoE	PE YoE	PE %	Adv. CAT YoE	MT good/bad for:	PE comment
ita-nohigh-fast	FL (FT)	C1	2-5	2-5	100%	Often	G: Productivity, quality, repetitive work.	PE better than from scratch when consistency is needed.
ita-nohigh-avg	FL (PT)	C1	>10	<2	20%	Often	G: Productivity, repetitive work. B: less creative.	PE produces unnatural sentences.
ita-nohigh-slow	FL (PT)	C2	>10	2-5	40%	Sometimes	G: creativity.	Good for time saving.
ita-oracle-fast	FL (FT)	C2	5-10	2-5	60%	Sometimes	G: Productivity, repetitive work. B: less creative.	Good for productivity, humans always needed.
ita-oracle-avg	FL (FT)	C2	5-10	5-10	20%	Always	G: productivity, terminology.	Good for tech docs, not for articulated texts.
ita-oracle-slow	FL (FT)	C2	2-5	5-10	80%	Always	G: Productivity, repetitive work.	Useful for consistency and productivity, unless creativity is needed.
ita-unsup-fast	FL (FT)	C1	<2	<2	60%	Often	G: Productivity, terminology. B: less creative.	Humans will always be needed in translation.
ita-unsup-avg	FL (FT)	C1	>10	2-5	60%	Often	G: Productivity, repetitive work. B: less creative.	An opportunity for translators.
ita-unsup-slow	FL (FT)	C1	5-10	5-10	80%	Always	G: Productivity, repetitive work. B: less creative.	Good for focusing on detailed/cultural/creative aspects of translations.
ita-sup-fast	FL (PT)	C1	>10	2-5	40%	Often	G: Productivity, quality, repetitive work, terminology.	Improves quality and consistency.
ita-sup-avg	FL (FT)	C1	>10	5-10	100%	Always	G: Productivity, repetitive work. B: less creative.	Consistency improved, but less variance means less creativity.
ita-sup-slow	FL (FT)	C1	>10	2-5	20%	Always	G: Productivity, creativity, quality, repetitive work.	Good for productivity, but does not work on creative texts.
nld-nohigh-fast	FL (FT)	C1	>10	>10	40%	Often	G: Productivity, terminology. B: creativity.	Widespread but still too literal
nld-nohigh-avg	FL (FT)	C2	>10	2-5	40%	Always	G: Repetitive work. B: creativity, often wrong, worse quality.	Increase in productivity to save on costs brings down quality.
nld-nohigh-slow	FL (FT)	C2	>10	5-10	100%	Often	G: Creativity, quality, repetitive work, terminology.	Working with MT can be creative beyond PE.
nld-oracle-fast	FL (FT)	C1	5-10	5-10	80%	Always	G: Productivity, quality, repetitive work, terminology.	Good for tech docs and repetition.
nld-oracle-avg	FL (FT)	C2	>10	2-5	40%	Always	B: less creative, less productive, often wrong	Bad MT is worse than no MT for specialized domains.
nld-oracle-slow	FL (FT)	C2	>10	2-5	60%	Often	G: Productivity, repetitive work. B: cultural references.	More productivity at the cost of idioms and cultural factors.
nld-unsup-fast	FL (FT)	C2	5-10	2-5	40%	Often	G: all. B: often wrong, worse quality.	PE makes you less in touch with the texts and often poorly paid.
nld-unsup-avg	FL (FT)	C2	5-10	2-5	60%	Sometimes	G: Productivity, quality, repetitive work, terminology. B: wrong.	Practical but less effective for longer passages.
nld-unsup-slow	FL (FT)	C2	>10	2-5	40%	Always	G: repetitive work, productivity, terminology	Improves consistency and productivity if applied well.
nld-sup-fast	FL (FT)	C2	>10	5-10	60%	Often	G: repetitive work, creativity, terminology	Useful, but worries about job loss
nld-sup-avg	FL (FT)	C2	>10	10	60%	Sometimes	G: terminology, creativity	Useful for inspiration on better translations
nld-sup-slow	FL (FT)	C1	5-10	5-10	80%	Always	G: repetitive work, productivity	Better productivity at the cost of creativity.

**Table C.8 |** Sample of pre-task questionnaire results. YoE = years of experience. FL = Freelance, PT = Part-time, FT = Full-time. PE = Post-editing. G = Good, B = Bad.

ID	Freq. Issues	Qual.	Flue.	Acc.	High. acc.	High. useful	Inter. clear	Hard?	Highlights statements					
									↑ Speed?	↑ Qual.?	↑ Ef- fort?	↑ Infl.	Spot errors	↑ En- joy?
ita-nohigh- fast	inflection, addi- tions, omissions	4	0.8	0.8	-	-	5	1	-	-	-	-	-	-
ita-nohigh- avg	multiple	3	0.6	0.4	-	-	2	4	-	-	-	-	-	-
ita-nohigh- slow	terminology, omissions	3	0.8	0.8	-	-	1	5	-	-	-	-	-	-
ita-oracle- fast	inflection, termi- nology	5	0.4	0.8	4	4	4	5	5	2	1	1	1	4
ita-oracle- avg	syntax, terminol- ogy, omissions, no context	3	0.4	0.6	2	1	2	3	1	1	4	1	1	1
ita-oracle- slow	syntax, no context	3	0.6	0.6	2	2	2	5	1	1	1	1	4	1
ita-unsup- fast	omissions	3	0.8	0.6	3	2	4	5	3	3	3	2	2	2
ita-unsup- avg	syntax, terminol- ogy, no context	3	0.6	0.6	3	3	3	5	2	3	2	1	1	3
ita-unsup- slow	syntax, inflection, terminology, omissions	3	0.4	0.6	2	2	3	4	2	2	3	3	4	4
ita-sup- fast	syntax, terminol- ogy, no context	3	0.4	0.4	2	1	2	2	1	1	3	1	2	2
ita-sup- avg	syntax, terminol- ogy, no context	3	0.4	0.4	2	2	3	5	3	2	4	3	3	4
ita-sup- slow	syntax, terminol- ogy, omissions, no context	3	0.6	0.6	2	2	1	2	2	1	1	4	4	1
nld-nohigh- fast	syntax, terminol- ogy, omissions, no context	3	0.2	0.4	-	-	4	4	-	-	-	-	-	-
nld-nohigh- avg	syntax, terminol- ogy, omissions, no context	2	0.4	0.6	-	-	4	5	-	-	-	-	-	-
nld-nohigh- slow	terminology, omissions, no context	2	0.2	0.4	-	-	3	5	-	-	-	-	-	-
nld-oracle- fast	syntax, inflection, terminology	3	0.6	0.6	2	1	3	2	2	2	2	1	1	1
nld-oracle- avg	syntax, terminol- ogy	3	0.8	0.6	4	3	3	4	3	3	3	3	2	3
nld-oracle- slow	syntax, terminol- ogy	3	0.6	0.4	3	1	3	4	1	1	1	1	1	3
nld-unsup- fast	terminology, addi- tions, omissions	3	0.6	0.8	3	2	4	4	1	3	1	1	2	1
nld-unsup- avg	multiple	3	0.6	0.6	4	3	2	4	3	3	4	3	2	3
nld-unsup- slow	syntax, terminol- ogy, omissions	1	0.4	0.4	2	4	1	4	4	4	3	2	2	3
nld-sup- fast	terminology, omissions, no context	3	0.6	0.4	2	2	3	5	1	1	5	3	1	1
nld-sup- avg	syntax, additions, no context	3	0.4	0.6	2	2	2	4	1	1	1	1	2	3
nld-sup- slow	multiple	5	0.8	1	4	3	2	5	3	3	2	2	2	4

**Table C.9 |** Sample of post-task questionnaire results. Statements use a 1–Strongly disagree to 5–Strongly agree scale.

<b>Target:</b> Seg. Edit Time, 5s bins from 0 to 600s		
Feature	Coeff.	Sig.
(Intercept)	1.67	***
MT Num. Chars	2.42	***
High. Ratio %	1.59	***
Target Lang.: ITA	-0.34	***
Domain: Social	0.31	***
<b>Oracle</b> Highlight	-0.79	.
<b>Sup.</b> Highlight	0.02	
<b>Unsup.</b> Highlight	-0.07	
MT xCOMET QE	0.01	***
ITA: <b>Oracle</b>	0.91	***
ITA: <b>Sup.</b>	1.18	***
ITA: <b>Unsup.</b>	0.48	***
Social: <b>Oracle</b>	-0.19	**
Social: <b>Sup.</b>	-0.34	***
Social: <b>Unsup.</b>	-0.22	***
High. Ratio: <b>Oracle</b>	-0.83	*
High. Ratio: <b>Sup.</b>	-1.33	***
Edit Order		
Translator ID		<b>Random Factors</b>
Segment ID		

**Table C.10** | Details for the negative binomial mixed-effect model used for the productivity analysis of Section 9.3.1.

<b>Target:</b> % edited char. in segment (0-100).		
Feature	Coeff.	Sig.
(Intercept)	21.0	***
MT Num. Chars	10.3	***
High. Ratio %	7.1	***
Target Lang.: ITA	-9.9	***
Domain: Social	10.9	***
<b>Oracle</b> Highlight	-5.2	
<b>Sup.</b> Highlight	-4.7	
<b>Unsup.</b> Highlight	-0.9	
ITA: <b>Oracle</b>	12.2	***
ITA: <b>Sup.</b>	15.9	***
ITA: <b>Unsup.</b>	13.4	***
Social: <b>Oracle</b>	3.5	***
Social: <b>Sup.</b>	-0.4	
Social: <b>Unsup.</b>	2.1	**
High. Ratio: <b>Oracle</b>	-0.18	
High. Ratio: <b>Sup.</b>	-1.78	***
Edit Order		
Translator ID		<b>Random Factors</b>
Segment ID		
MT # Chars		
Target Lang		<b>Zero-Inflation Factors</b>
Text Domain		
Translator ID		

**Table C.11** | Details for the zero-inflated negative binomial mixed-effect model used for the editing analysis of Section 9.3.2. The model achieves an RMSE of 0.11 and an  $R^2$  of 0.98.

Modalities	English→Italian			English→Dutch			Both		
	Bio	Social	Both	Bio	Social	Both	Bio	Social	Both
Oracle and <b>Sup.</b>	0.17	0.32	0.25	<b>0.38</b>	0.29	0.34	0.26	0.29	0.29
Oracle and <b>Unsup.</b>	0.14	0.30	0.20	<b>0.31</b>	0.27	0.28	0.22	0.29	0.24
Supervised and <b>Oracle</b>	0.19	<b>0.31</b>	0.26	0.30	0.26	0.29	0.24	0.29	0.28
Supervised and <b>Unsup.</b>	0.19	<b>0.33</b>	0.25	0.28	0.24	0.25	0.24	0.29	0.25
Unsupervised and <b>Oracle</b>	0.22	0.32	0.27	<b>0.35</b>	0.30	0.33	0.28	0.31	0.30
Unsupervised and <b>Sup.</b>	0.22	0.37	0.30	<b>0.39</b>	0.27	0.33	0.30	0.31	0.32

**Table C.12 |** Average highlight agreement proportion between different modalities across language pairs and domains (Section 9.3.2). Scores are normalized to account for the relative frequency of highlight modalities compared to the mean highlight frequency for the current language and domain combination.

Domain	Speed	$P(H)$	$P(E)$	$P(E H)$	$P(E \neg H)$	$\Lambda_H(E)$	$P(H E)$	$P(H \neg E)$	$\Lambda_E(H)$
<b>English→Italian</b>									
Biomed.	Fast	.04/.01	.12/.02	.03/.01	4.0/2.0	.30/.27	.08/.11	3.7/2.4	
	Avg.	.09	.10/.05	.27/.12	.09/.04	3.0/3.0	.22/.30	.07/.11	3.1/2.7
	Slow	.09/.02	.21/.04	.08/.01	2.6/4.0	.19/.26	.07/.11	2.7/2.3	
Social	Fast	.11/.07	.30/.20	.07/.04	4.2/5.0	.40/.52	.11/.16	3.6/3.2	
	Avg.	.14	.23/.14	.48/.32	.18/.10	2.6/3.2	.30/.42	.09/.15	3.3/2.8
	Slow	.17/.05	.39/.14	.14/.03	2.7/4.6	.31/.54	.11/.17	2.8/3.1	
<b>English→Dutch</b>									
Biomed.	Fast	.03/.02	.11/.05	.02/.01	5.5/5.0	.48/.61	.13/.18	3.6/3.3	
	Avg.	.14	.11/.19	.20/.30	.10/.17	2.0/1.7	.25/.29	.13/.16	1.9/1.8
	Slow	.12/.10	.26/.23	.10/.07	2.6/3.2	.29/.42	.12/.16	2.4/2.6	
Social	Fast	.06/.07	.19/.21	.04/.04	4.7/5.2	.37/.47	.10/.13	3.7/3.6	
	Avg.	.12	.17/.32	.32/.48	.15/.29	2.1/1.6	.22/.23	.10/.12	2.2/1.9
	Slow	.18/.18	.38/.40	.15/.14	2.5/2.8	.25/.34	.09/.11	2.7/3.0	

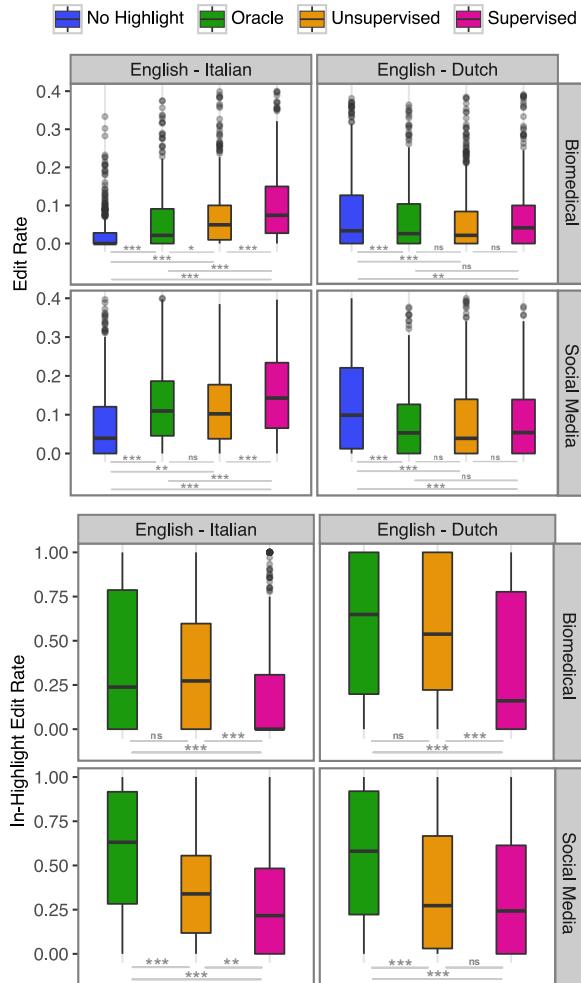
**Table C.13 |** Highlighting ( $H$ ) and editing ( $E$ ) statistics for each domain, and translation direction across translator speeds ( $n = 4$  post-editors per combination, regardless of highlight modality). Values after slashes are adjusted by projecting highlights of the specified modality over edits from **No Highlight** translators to estimate highlight-induced editing biases (Section 9.3.2).

Domain	Modality	$P(H)$	$P(E)$	$P(E H)$	$P(E \neg H)$	$\Lambda_H(E)$	$P(H E)$	$P(H \neg E)$	$\Lambda_E(H)$
<b>English → Italian</b>									
Biomed.	<b>Random</b>	.12	-	-.02	-.02	-.1/0	-.1/11	-.1/13	-.1/0.8
	<b>No High.</b>	-	.02	-	-	-	-	-	-
	<b>Oracle</b>	.08	.07	.26/.08	.05/.02	5.2/4.0	.30/.26	.06/.08	5.0/3.2
	<b>Unsup.</b>	.16	.10	.18/.06	.08/.02	2.2/3.0	.29/.36	.14/.15	2.0/2.4
	<b>Sup.</b>	.11	.12	.18/.05	.11/.02	1.6/2.5	.16/.23	.10/.10	1.6/2.3
Social	<b>Random</b>	.20	-	-.09	-.09	-.1/0	-.1/21	-.1/20	-.1/1.0
	<b>No High.</b>	-	.09	-	-	-	-	-	-
	<b>Oracle</b>	.25	.20	.42/.23	.13/.04	3.2/5.7	.52/.66	.18/.21	2.8/3.1
	<b>Unsup.</b>	.17	.18	.35/.19	.14/.07	2.5/2.7	.33/.37	.14/.15	2.3/2.4
	<b>Sup.</b>	.15	.21	.38/.23	.18/.06	2.1/3.8	.27/.39	.11/.12	2.4/3.2
<b>English → Dutch</b>									
Biomed.	<b>Random</b>	.17	-	-.12	-.10	-.1/2	-.1/19	-.1/17	-.1/1.1
	<b>No High.</b>	-	.10	-	-	-	-	-	-
	<b>Oracle</b>	.21	.08	.21/.20	.05/.08	4.2/2.5	.52/.41	.18/.18	2.8/2.2
	<b>Unsup.</b>	.23	.09	.17/.17	.07/.08	2.4/2.1	.43/.38	.21/.21	2.0/1.8
	<b>Sup.</b>	.12	.08	.20/.21	.06/.09	3.3/2.3	.30/.25	.11/.11	2.7/2.2
Social	<b>Random</b>	.16	-	-.22	-.19	-.1/1	-.1/19	-.1/16	-.1/1.1
	<b>No High.</b>	-	.19	-	-	-	-	-	-
	<b>Oracle</b>	.19	.12	.33/.39	.07/.15	4.7/2.6	.54/.39	.15/.15	3.6/2.6
	<b>Unsup.</b>	.15	.13	.25/.33	.11/.17	2.2/1.9	.30/.26	.13/.12	2.3/2.1
	<b>Sup.</b>	.12	.10	.30/.36	.08/.17	3.7/2.1	.36/.23	.10/.10	3.6/2.3

**Table C.14 |** Highlighting ( $H$ ) and editing ( $E$ ) statistics for each domain, modality and translation direction combination ( $n = 3$  post-editors per combination). Values after slashes are adjusted by projecting highlights of the specified modality over edits from No HIGHLIGHT translators to estimate highlight-induced editing biases (Section 9.3.2). A Random baseline is added by projecting random highlights matching the average frequency over all modalities for specific domain and translation direction settings.

ID	Source text	Target text	Proposed correction	Error Annotation		Score
				Category	Severity	
9-1	Specifying peri- and postnatal factors in children born very preterm (VPT) that affect later outcome helps to improve long-term treatment.	Specificare i fattori peri- e postnatali nei bambini nati molto pretermine (VPT) che influenzano il risultato successivo aiuta a migliorare il trattamento a lungo termine.	Specificare i fattori peri- e postnatali nei bambini nati molto pretermine (VPT, Very Preterm) che influenzano il risultato successivo aiuta a migliorare il trattamento a lungo termine.	Readability	Minor	90
9-2	To enhance the predictability of 5-year cognitive outcome by perinatal, 2-year developmental and socio-economic data.	Migliorare la prevedibilità del risultato cognitivo a 5 anni mediante dati perinatali, di sviluppo e socioeconomici a 2 anni.				100
9-3	5-year infants born VPT were compared to 34 term controls.	I neonati di 5 anni nati VPT sono stati confrontati con 34 <b>nati a termine come controllo</b> .	I neonati di 5 anni nati VPT sono stati confrontati con 34 <b>controlli a termine</b> .	Mistranslation	Minor	70
9-4	The IQ of 5-year infants born VPT was 10 points lower than that of term controls and influenced independently by preterm birth and SES.	Il IQ dei bambini di 5 anni nati VPT era di 10 punti inferiore a quello dei nati a termine <b>di controllo</b> , e influenzato indipendentemente dalla nascita pretermine e dai dati SES.	Il IQ dei bambini di 5 anni nati VPT era di 10 punti inferiore a quello dei nati a termine e influenzato indipendentemente dalla nascita pretermine e dallo stato socioeconomico (SES).	Mistranslation	Minor	70
		Il IQ dei bambini di 5 anni nati VPT era di 10 punti inferiore a quello dei nati a termine di controllo, e influenzato indipendentemente dalla nascita pretermine e <b>dai dati SES</b> .	Il IQ dei bambini di 5 anni nati VPT era di 10 punti inferiore a quello dei nati a termine e influenzato indipendentemente dalla nascita pretermine e <b>dallo stato socioeconomico (SES)</b> .	Untranslated	Minor	
52-1	But with less than 3 months to go for that, I feel I'm not ready yet, but having never taken it, I have nothing to compare it to besides colleagues' advice.	Ma con meno di 3 mesi per farlo, sento di non essere ancora pronto, <b>ma non l'ho mai preso</b> , non ho nulla con cui confrontarlo oltre ai consigli dei colleghi.	Ma con meno di 3 mesi per farlo, sento di non essere ancora pronto, <b>e non avendolo mai fatto</b> , non ho nulla con cui confrontarlo oltre ai consigli dei colleghi.	Mistranslation	Major	30
52-2	Without knowing what I know, they can't know if I'm actually ready yet, but many of them are pushing me to sign up for it.	<b>Senza sapere quello che so</b> , non possono sapere se sono ancora pronta, ma molti di loro mi stanno spingendo a iscrivermi.	<b>Se non hanno idea di quanto sappia</b> , non possono sapere se sono davvero pronta, ma molti di loro mi stanno spingendo a iscrivermi.	Readability	Minor	60
		Senza sapere quello che so, non possono sapere se sono <b>ancora</b> pronta, ma molti di loro mi stanno spingendo a iscrivermi.	Se non hanno idea di quanto sappia, non possono sapere se sono <b>davvero</b> pronta, ma molti di loro mi stanno spingendo a iscrivermi.	Mistranslation	Minor	
52-3	I'm close... I just don't know if I'm 2 months study close.	Ci sono quasi... solo che non so se ce la farò in soli 2 mesi, <b>ma penso di potercela fare</b> .	Ci sono quasi... solo che non so se ce la farò in soli 2 mesi.	Addition	Major	20

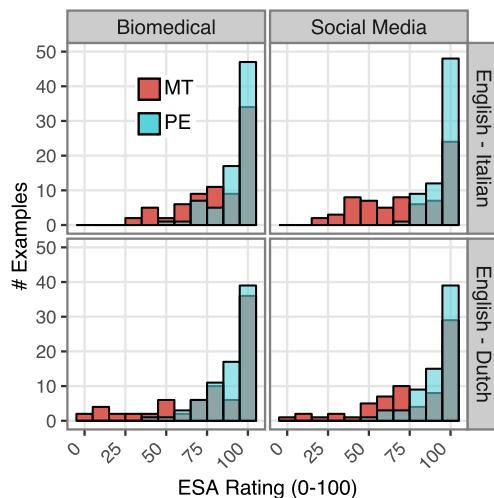
**Table C.15 | QA interface with cropped examples of biomedical and social media texts with error annotations (Biomedical: post-edited segments with No HIGHLIGHT; Social media: MT outputs).**



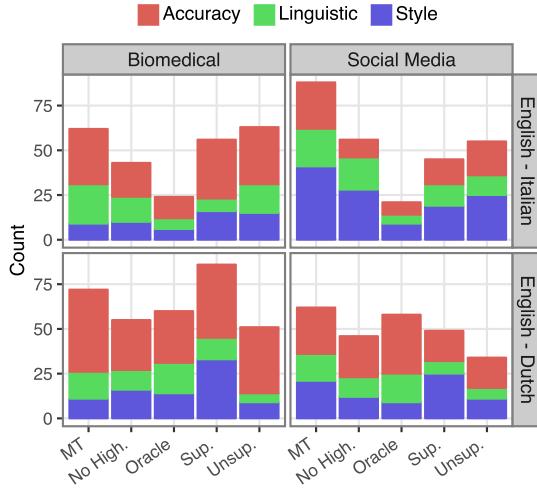
**Figure C.5 | Top:** Post-editing rate across highlight modalities, domains and directions.  
**Bottom:** Proportion of edits in highlighted spans across highlight modalities. \*\*\* =  $p < 0.001$ , \*\* =  $p < 0.01$ , \* =  $p < 0.05$ , ns = not significant with Bonferroni correction.

		Do post-edits in QE: en → it			en → nl		
		Oracle	Sup'd	Unsup'd	Oracle	Sup'd	Unsup'd
Also happen in:	Oracle	79%	75%	66%	72%	75%	67%
	Supervised	82%	82%	73%	78%	82%	72%
	Unsupervised	76%	76%	71%	75%	78%	70%
	No highlight	60%	62%	52%	79%	82%	75%

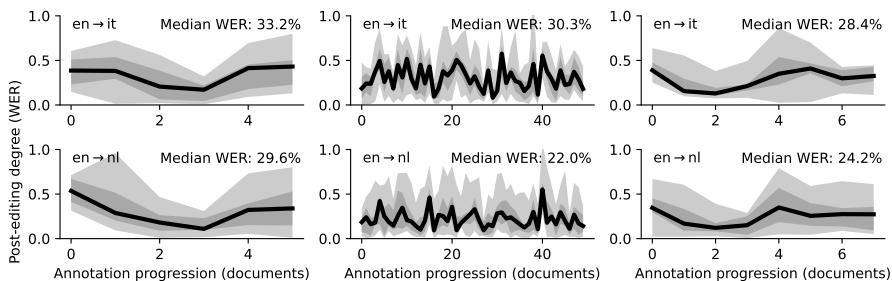
**Figure C.6** | Post-editing agreement across various modalities Section 9.3.2. Results are averaged across all translator pairs for the two modalities ( $n = 3$  intra-modality,  $n = 9$  inter-modality for every language) and all segments.



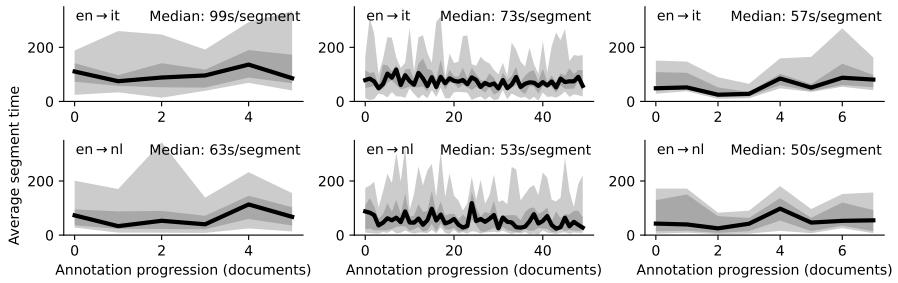
**Figure C.7** | ESA ratings for MT outputs and post-edits across domains and translation directions.



**Figure C.8 |** Distribution of MQM error categories for MT and post-edits across highlight modalities for the two translation directions and domains of QE4PE.



**Figure C.9 |** Editing proportion, measured by word error rate between MT and post-edited texts, with respect to post-editor progression. Values are medians across all post-editors.



**Figure C.10 | Segment-level post-editing time with respect to post-editor progression.** Values are medians across all annotators. Light gray area is min-max values, dark gray represents 25%-75% quantiles. The annotators do not become considerably faster with the task progression, likely due to the simplicity of the task and the high post-editing proficiency of professional post-editors. The high variability in editing times motivates the careful group assignments performed using PRE task edit logs.

### C.3 Unsupervised MT Error Detection and Human Disagreement

#### C.3.1 Full Results

Method	QE4PE <sub>t1</sub>		QE4PE <sub>t2</sub>		QE4PE <sub>t3</sub>		QE4PE <sub>t4</sub>		QE4PE <sub>t5</sub>		QE4PE <sub>t6</sub>		QE4PE <sub>avg</sub>	
	AP	F1*	AP	F1*	AP	F1*	AP	F1*	AP	F1*	AP	F1*	AP	F1*
Random Baseline	.08	.14	.15	.26	.06	.12	.11	.19	.22	.36	.18	.30		.13 .23
Surprisal	.11	.20	.21	.31	.11	.17	.16	.25	.30	.40	.25	.35		.19 .28
Out. Entropy	.12	.18	.22	.30	.10	.16	.17	.24	.30	.39	.26	.34		.19 .27
Surprisal MCD <sub>AVG</sub>	.12	.20	.22	.32	.11	.17	.16	.26	.30	.41	.26	.36		.19 .29
Surprisal MCD <sub>VAR</sub>	.13	.21	.26	.33	.12	.20	.19	.27	.31	.40	.29	.36		.22 .30
LL Surprisal <sub>BEST</sub>	.11	.19	.21	.32	.11	.16	.16	.25	.29	.40	.26	.35		.19 .28
LL KL-Div <sub>BEST</sub>	.09	.16	.19	.28	.08	.14	.13	.21	.25	.37	.22	.31		.16 .25
LL Pred. Depth	.09	.16	.18	.28	.07	.13	.14	.21	.25	.37	.21	.31		.16 .24
Attn. Entropy <sub>AVG</sub>	.11	.16	.17	.27	.12	.17	.11	.19	.23	.36	.19	.31		.15 .24
Attn. Entropy <sub>MAX</sub>	.09	.14	.15	.26	.10	.18	.09	.19	.20	.36	.16	.30		.13 .24
BLOOD <sub>BEST</sub>	.08	.14	.16	.26	.06	.12	.11	.19	.23	.36	.18	.30		.14 .23
XCOMET-XL	.11	.24	.22	.35	.10	.20	.16	.30	.27	.35	.23	.34		.18 .30
XCOMET-XL <sub>CONF</sub>	.20	.25	.30	.36	.14	.21	.25	.31	.37	.40	.31	.36		.26 .32
XCOMET-XXL	.13	.27	.22	.32	.10	.24	.17	.31	.28	.32	.23	.31		.19 .30
XCOMET-XXL <sub>CONF</sub>	.19	.27	.31	.36	.17	.24	.26	.32	.37	.41	.33	.39		.27 .33
Human Editors <sub>MIN</sub>	.17	.33	.26	.38	.10	.21	.16	.26	.25	.36	.23	.30		.19 .31
Human Editors <sub>AVG</sub>	.20	.38	.29	.43	.14	.30	.22	.39	.32	.38	.30	.40		.25 .39
Human Editors <sub>MAX</sub>	.24	.43	.31	.47	.20	.41	.24	.43	.37	.50	.33	.50		.28 .46

**Table C.16 |** WQE metrics' performance for predicting error spans from the six edit sets over NLLB 3.3B translations in the En→It QE4PE dataset. Best unsupervised and overall best metric results are highlighted.

Method	QE4PE <sub>t1</sub>		QE4PE <sub>t2</sub>		QE4PE <sub>t3</sub>		QE4PE <sub>t4</sub>		QE4PE <sub>t5</sub>		QE4PE <sub>t6</sub>		QE4PE <sub>avg</sub>	
	AP	F1*	AP	F1*	AP	F1*	AP	F1*	AP	F1*	AP	F1*	AP	F1*
Random Baseline	.07	.14	.34	.51	.22	.36	.19	.32	.13	.24	.22	.36	.20	.32
Surprisal	.12	.19	.41	.51	.30	.39	.29	.37	.21	.30	.31	.41	.27	.36
Out. Entropy	.11	.18	.41	.51	.31	.37	.29	.36	.20	.27	.31	.39	.27	.35
Surprisal <sub>MCD AVG</sub>	.12	.19	.42	.52	.31	.40	.30	.40	.21	.30	.31	.42	.28	.37
Surprisal <sub>MCD VAR</sub>	.13	.21	.45	.53	.36	.41	.34	.40	.24	.32	.36	.42	.31	.38
LL Surprisal <sub>BEST</sub>	.12	.19	.42	.53	.30	.40	.29	.38	.21	.30	.31	.41	.27	.37
LL KL-Div <sub>BEST</sub>	.09	.15	.39	.52	.28	.37	.25	.34	.17	.26	.29	.38	.25	.34
LL Pred. Depth	.09	.16	.37	.52	.26	.37	.24	.33	.17	.25	.27	.38	.23	.33
Attn. Entropy <sub>Avg</sub>	.09	.15	.37	.51	.22	.36	.20	.32	.13	.24	.23	.37	.21	.32
Attn. Entropy <sub>Max</sub>	.09	.15	.35	.51	.22	.36	.18	.32	.12	.24	.21	.37	.19	.32
BLOOD <sub>BEST</sub>	.07	.13	.35	.51	.22	.36	.19	.32	.14	.24	.23	.36	.20	.32
XCOMET-XL	.13	.27	.39	.39	.31	.44	.28	.32	.20	.35	.31	.44	.27	.38
XCOMET-XL <sub>CONF</sub>	.24	.31	.47	.53	.43	.45	.40	.43	.29	.36	.43	.46	.38	.42
XCOMET-XXL	.13	.28	.39	.29	.30	.35	.26	.35	.19	.31	.30	.35	.26	.32
XCOMET-XXL <sub>CONF</sub>	.24	.30	.48	.53	.43	.45	.40	.42	.31	.35	.43	.45	.38	.42
Human Editors <sub>MIN</sub>	.16	.29	.43	.51	.34	.45	.33	.47	.26	.42	.36	.46	.32	.43
Human Editors <sub>Avg</sub>	.17	.33	.44	.51	.34	.45	.33	.47	.26	.42	.36	.46	.32	.43
Human Editors <sub>MAX</sub>	.19	.36	.46	.51	.36	.51	.37	.53	.32	.51	.40	.53	.35	.49

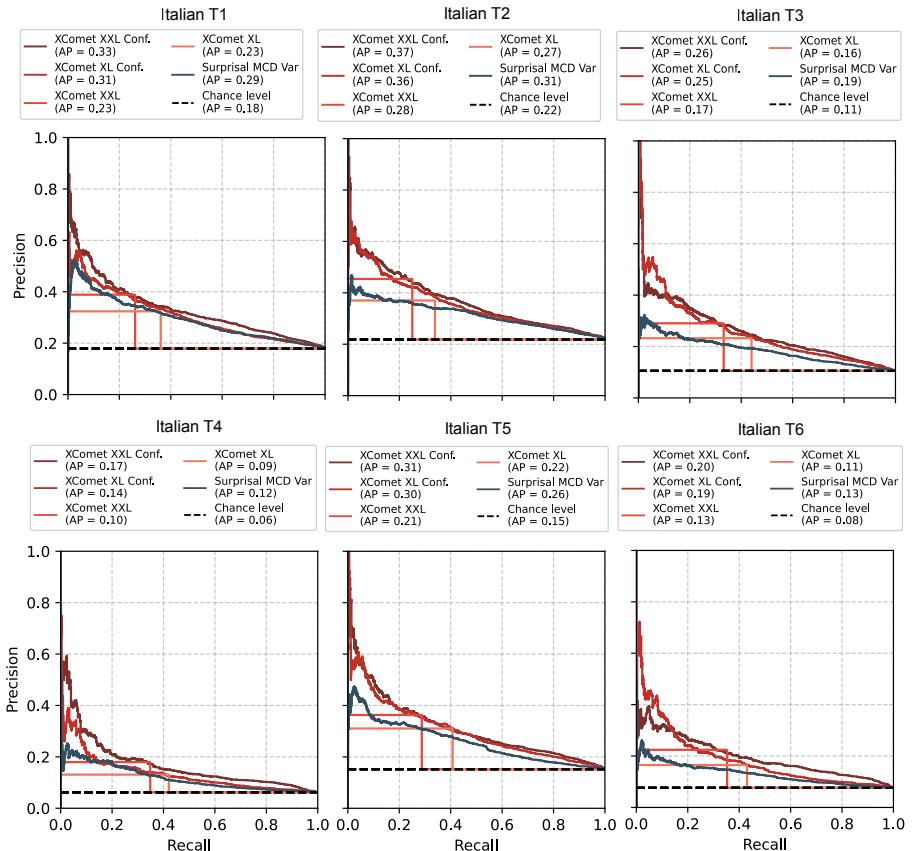
**Table C.17 |** WQE metrics’ performance for predicting error spans from the six edit sets over NLLB 3.3B translations in the En→Ni QE4PE dataset. Best unsupervised and overall best metric results are highlighted.

Method	Italian		Dutch		Arabic		Turkish		Vietnamese		Ukrainian		Average	
	AP	F1*	AP	F1*	AP	F1*	AP	F1*	AP	F1*	AP	F1*	AP	F1*
Random Baseline	.25	.40	.28	.43	.33	.49	.34	.50	.35	.52	.48	.65	.34	.50
Surprisal	.34	.45	.36	.46	.42	.51	.43	.54	.46	.55	.55	.65	.43	.53
Out. Entropy	.37	.43	.39	.45	.45	.50	.49	.52	.48	.54	.58	.65	.46	.51
Surprisal <sub>MCD AVG</sub>	.34	.45	.37	.47	.43	.52	.44	.54	.46	.55	.56	.65	.43	.53
Surprisal <sub>MCD VAR</sub>	.39	.46	.41	.47	.47	.53	.49	.55	.48	.55	.61	.67	.48	.54
LL Surprisal <sub>BEST</sub>	.33	.44	.36	.45	.41	.51	.44	.54	.44	.55	.55	.66	.42	.53
LL KL-Div <sub>BEST</sub>	.34	.42	.37	.45	.41	.51	.44	.52	.44	.52	.56	.65	.43	.51
LL Pred. Depth	.30	.42	.32	.44	.39	.50	.40	.52	.39	.53	.54	.66	.39	.51
Attn. Entropy <sub>Avg</sub>	.28	.41	.30	.43	.35	.49	.37	.51	.40	.52	.50	.65	.37	.50
Attn. Entropy <sub>Max</sub>	.25	.41	.26	.43	.34	.49	.34	.50	.35	.52	.47	.65	.34	.50
BLOOD <sub>BEST</sub>	.26	.40	.28	.43	.35	.52	.35	.50	.36	.52	.49	.65	.35	.51
XCOMET-XL	.34	.39	.37	.44	.41	.47	.44	.50	.42	.44	.56	.44	.42	.45
XCOMET-XL <sub>CONF</sub>	.46	.47	.49	.50	.51	.53	.58	.56	.53	.55	.68	.67	.54	.55
XCOMET-XXL	.34	.36	.35	.35	.43	.47	.45	.48	.43	.42	.57	.41	.43	.42
XCOMET-XXL <sub>CONF</sub>	.48	.49	.50	.50	.55	.54	.58	.56	.56	.57	.70	.67	.56	.55

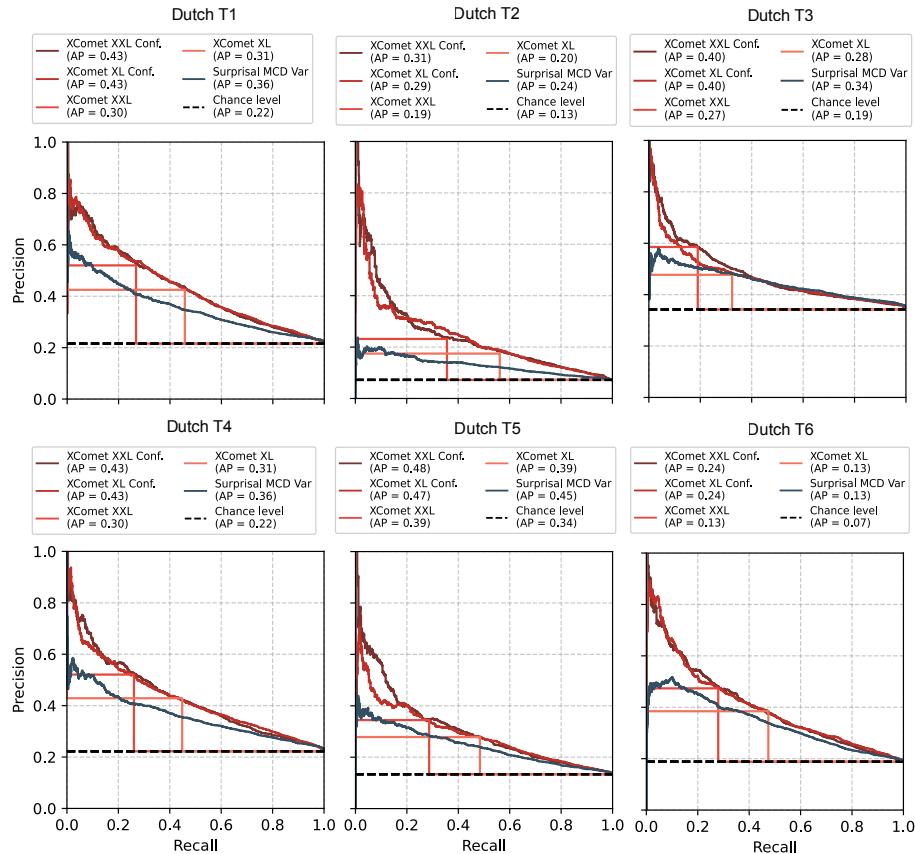
**Table C.18 |** WQE metrics’ performance for predicting error spans from multiple edit sets (one per language) over mBART-50 translations across the six topologically diverse target languages of DivEMT.

Method	En→Ja		En→Zh		En→Hi		Cs→Uk		En→Cs		En→Ru		Average	
	AP	F1*	AP	F1*	AP	F1*	AP	F1*	AP	F1*	AP	F1*	AP	F1*
Random Baseline	.02	.03	.03	.07	.03	.07	.05	.09	.06	.11	.08	.16	.05	.09
Surprisal	<u>.03</u>	.07	.05	.09	.05	.09	.14	.20	.10	.16	.13	.19	.08	.13
Out. Entropy	<u>.03</u>	<u>.08</u>	.06	<u>.11</u>	.06	.10	<b>.20</b>	<u>.27</u>	<u>.12</u>	<u>.18</u>	<u>.14</u>	<u>.20</u>	<u>.10</u>	<u>.16</u>
LL Surprisal <sub>BEST</sub>	<u>.03</u>	.07	.05	.09	.05	.09	.14	.20	.10	.16	.13	.19	.08	.13
LL KL-Div <sub>BEST</sub>	.02	.05	.04	.07	.04	.08	.10	.17	.09	.15	.12	.19	.07	.12
LL Pred. Depth	.02	.05	.04	.08	.04	.09	.09	.18	.08	.14	.11	.18	.06	.12
Attn. Entropy <sub>AVG</sub>	.02	.03	.03	.07	.03	.07	.03	.09	.05	.11	.07	.16	.04	.09
Attn. Entropy <sub>MAX</sub>	.01	.03	.03	.07	.03	.07	.03	.09	.05	.11	.08	.16	.04	.09
XCOMET-XL	.04	.09	.05	.11	.06	.12	.13	.28	.11	.24	.16	.32	.09	.19
XCOMET-XL <sub>CONF</sub>	<b>.08</b>	.14	<b>.10</b>	.16	<b>.10</b>	<b>.19</b>	.18	<b>.30</b>	.19	.29	.24	.32	.15	.23
XCOMET-XXL	.04	.11	.06	.13	.05	.11	.13	.28	.11	.24	.16	<b>.33</b>	.09	.20
XCOMET-XXL <sub>CONF</sub>	.07	<b>.15</b>	.09	<b>.19</b>	.09	.17	.19	.29	<b>.22</b>	<b>.30</b>	<b>.28</b>	<b>.33</b>	<b>.16</b>	<b>.24</b>

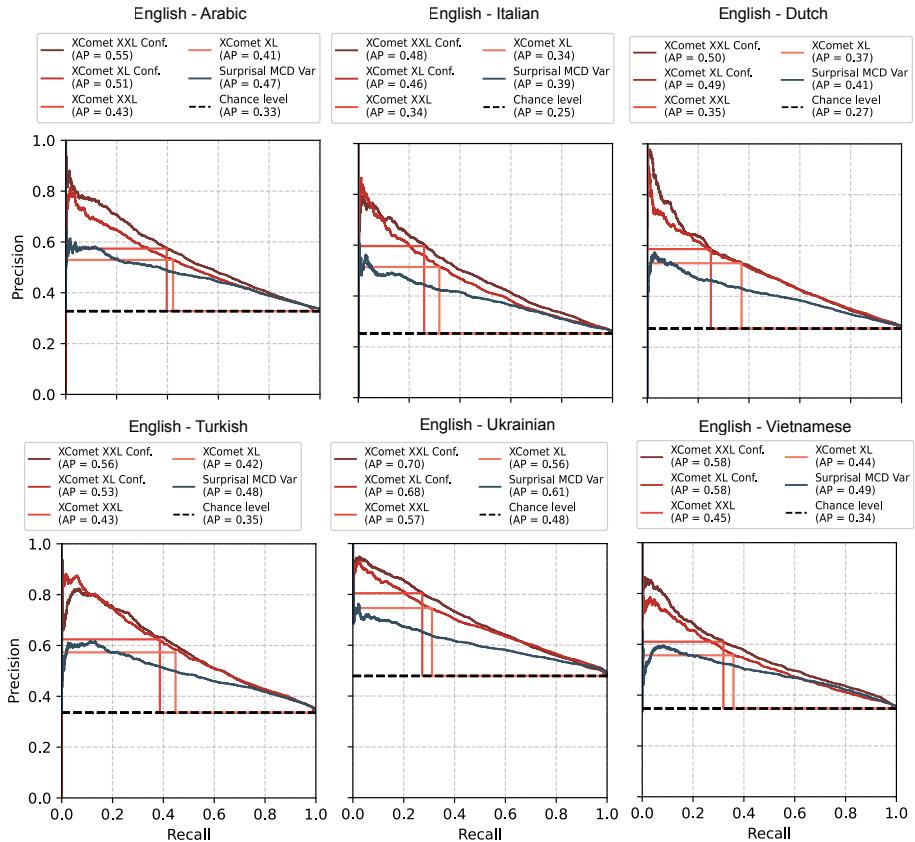
**Table C.19 |** WQE metrics' performance for predicting error spans from the ESA annotations (one set per language) over Aya23-35B outputs for the WMT24 dataset.



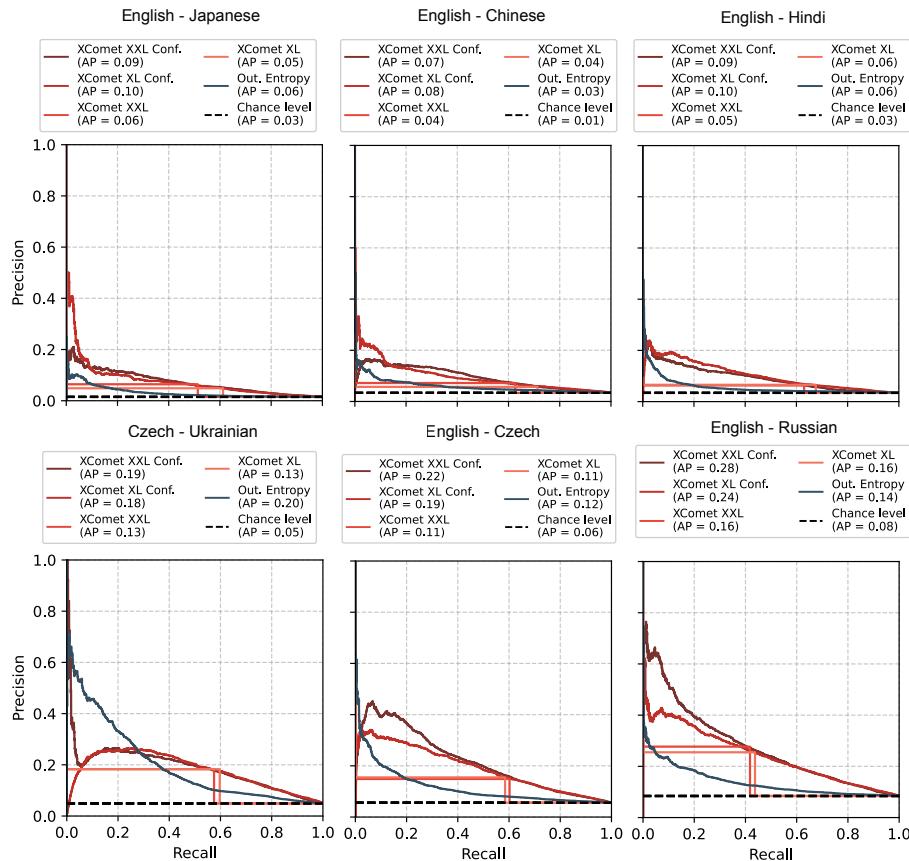
**Figure C.11 |** Precision-recall curves for xcomet metrics and Surprisal MCD<sub>var</sub> for all annotators of QE4PE En→It.



**Figure C.12 |** Precision-recall curves for xcomet metrics and Surprisal MCD<sub>var</sub> for all annotators of QE4PE En→Nl.



**Figure C.13 |** Precision-recall curves for xCOMET metrics and Surprisal MCD<sub>var</sub> on all DivEMT languages.



**Figure C.14 |** Precision-recall curves for xcomet metrics and Out. Entropy on all WMT24 languages.

# Summary

Neural language models have revolutionized the field of natural language processing, quickly becoming essential tools for a wide range of practical applications. Recent advances in interpretability research offered valuable insights into the inner workings of these systems, but often failed to translate into downstream improvements for users in real-world settings. This dissertation investigates the end-to-end development of interpretability methods to improve the trustworthiness and controllability of neural machine translation systems, from conception to experimentation with end users. Its findings address fundamental questions about how language models leverage contextual information, how their generation processes can be steered for personalization, and how interpretability insights can enhance professional translation practices.

The thesis work is organized into three interconnected parts. **Part I** develops foundational tools and methods for understanding how language models use contextual information during generation. We begin by introducing Inseq, an open-source toolkit for interactive analysis of language model behavior, showcasing its use for gender bias detection in machine translation and activation attribution using gradient-based methods. We then design PECoRE, a framework using contrastive input attribution to quantify how language models exploit contextual information, and demonstrate its effectiveness in detecting context influence in context-aware machine translation systems. Finally, we extend PECoRE to retrieval-augmented generation, using model internals to produce faithful, efficient and high-quality citations for open-book question answering.

**Part II** shifts the focus of our investigation from analysis to intervention, exploring methods for controlling translation outputs through prompting-based and steering-based approaches. We first present RAMP, a retrieval-augmented prompting technique exploiting relevant examples and style labels for attribute-controlled translation. We then move to the more challenging domain of literary translation, highlighting the effectiveness of steering interventions in conditioning models' generation by surgically altering their internal representations. In particular, we show that interpretable concepts extracted by trained sparse autoencoders can be used to mimic personal translation

styles from human professional translators, and that successful prompting and steering approaches converge on similar mechanistic solutions.

Finally, **Part III** explores how insights from model internals can inform human editors in professional translation workflows. We begin by conducting a post-editing user study spanning six typologically diverse languages (DivEMT), showing that translation productivity gains vary dramatically across language pairs, with typological similarity being more influential than traditional quality metrics. Our second study, QE4PE, investigates how word-level error highlights impact the productivity of professional post-editors and the quality of their translations, including both supervised and interpretability-based approaches. We conclude with a broad evaluation of unsupervised quality estimation methods, showing that error detection approaches based on model internals can outperform supervised baselines, and highlighting the importance of calibration and multiple annotations to account for human label variation.

Overall, this dissertation advances the field of machine translation interpretability by developing accessible tools and methods for understanding context usage, enabling fine-grained control over translation outputs, and establishing empirical evidence for the use of model internals in professional translation workflows. These contributions, taken together, lay the groundwork for the next generation of trustworthy, controllable, and user-centered translation systems.

# Samenvatting

Neurale taalmodellen hebben voor een revolutie gezorgd in het vakgebied van de natuurlijke taalverwerking en zijn snel uitgegroeid tot essentiële instrumenten voor een breed scala aan praktische toepassingen. Recente vorderingen in het onderzoek naar *interpreteerbaarheid* boden waardevolle inzichten in de interne werking van deze systemen, maar slaagden er vaak niet in zich te vertalen naar verbeteringen voor gebruikers in reële scenario's. Dit proefschrift onderzoekt de ontwikkeling van methoden, gebaseerd op interpreteerbaarheidsonderzoek, om de betrouwbaarheid en bestuurbaarheid van neurale machinevertaalsystemen te verbeteren, op een *end-to-end*-manier: van concept tot en met experimenteren met eindgebruikers. De bevindingen gaan in op fundamentele vragen over hoe taalmodellen contextuele informatie benutten, hoe hun generatieprocessen kunnen worden gestuurd voor personalisatie, en hoe inzichten uit interpreteerbaarheidsonderzoek de professionele vertaalpraktijk kunnen verbeteren.

Het proefschrift is opgebouwd in drie onderling verbonden delen. **Deel I** ontwikkelt fundamentele instrumenten en methoden om te begrijpen hoe taalmodellen contextuele informatie gebruiken tijdens het genereren van tekst. Om te beginnen introduceren we Inseq, een open-source toolkit voor interactieve analyse van taalmodelgedrag, en laten we zien hoe het gebruikt kan worden voor het detecteren van genderbias in machinevertaling en voor activatie-attributie met gradiënt-gbaseerde methoden. Vervolgens ontwerpen we PECoRE, een raamwerk dat *contrastieve inputattributie* gebruikt om te kwantificeren hoe taalmodellen contextuele informatie benutten, en we tonen de effectiviteit hiervan aan bij het detecteren van contextinvloed in contextbewuste machinevertaalsystemen. Tot slot breiden we PECoRE uit naar 'retrieval-augmented generation' (RAG), waarbij we de interne werking van het model gebruiken om betrouwbare, efficiënte en hoogwaardige citaten te produceren voor het beantwoorden van openboekvragen.

**Deel II** verschuift de focus van ons onderzoek van analyse naar interventie, en verkent methoden om de vertaaloutput te beheersen met technieken gebaseerd op *prompting* en *sturing*. We presenteren eerst RAMP, een 'retrieval-augmented prompting'-techniek die relevante voorbeelden en stijl-labels benut voor attribuut-gestuurde vertaling.

Vervolgens verplaatsen we ons naar het meer uitdagende domein van de literaire vertaling, waarbij we de effectiviteit benadrukken van sturingsinterventies bij het conditioneren van de modelgeneratie door hun interne representaties chirurgisch aan te passen. In het bijzonder tonen we aan dat interpreerbare concepten, geëxtraheerd door getrainde sparse autoencoders, kunnen worden gebruikt om persoonlijke vertaalstijlen van menselijke professionele vertalers na te bootsen, en dat succesvolle prompting- en sturing-benaderingen convergeren naar vergelijkbare mechanistische oplossingen.

Ten slotte onderzoekt **Deel III** hoe inzichten uit de interne werking van modellen menselijke post-editors kunnen informeren binnen professionele vertaalworkflows. We beginnen met het uitvoeren van een gebruikersstudie over post-editing in zes typologisch diverse talen omvat (DivEMT). Deze studie toont aan dat de productiviteitswinst bij vertalen dramatisch varieert tussen taalparen, waarbij typologische afstand invloedrijker is dan traditionele kwaliteitsmetrieken. Onze tweede studie, QE4PE, onderzoekt hoe foutmarkeringen op woordniveau de productiviteit van professionele post-editors en de kwaliteit van hun vertalingen beïnvloeden, waarbij zowel gesuperviseerde als op interpreerbaarheid gebaseerde benaderingen worden meegenomen. We sluiten af met een brede evaluatie van niet-gesuperviseerde methoden voor kwaliteitsbeoordeling, waaruit blijkt dat foutdetectiebenaderingen gebaseerd op interpreerbaarheid beter kunnen presteren dan gesuperviseerde baselines. We benadrukken tevens het belang van kalibratie en meervoudige annotaties om rekening te houden met menselijke variatie in labeling.

In het geheel genomen levert dit proefschrift een bijdrage aan het vakgebied van de interpreerbaarheid van machinevertaling door toegankelijke tools en methoden te ontwikkelen voor het begrijpen van contextgebruik, het mogelijk maken van fijnmazige beheersing van de vertaaloutput, en het leveren van empirisch bewijs voor het gebruik van interpreerbaarheid in professionele vertaalworkflows. Deze bijdragen leggen samen het fundament voor de volgende generatie van betrouwbare, beheersbare en gebruikersgerichte vertalsystemen.

## Groningen Dissertations in Linguistics (GRODIL)

---

1. Henriëtte de Swart (1991). Adverbs of Quantification: A Generalized Quantifier Approach.
2. Eric Hoekstra (1991). Licensing Conditions on Phrase Structure.
3. Dicky Gilbers (1992). Phonological Networks. A Theory of Segment Representation.
4. Helen de Hoop (1992). *Case Configuration and Noun Phrase Interpretation*.
5. Gosse Bouma (1993). *Nonmonotonicity and Categorial Unification Grammar*.
6. Peter I. Blok (1993). *The Interpretation of Focus*.
7. Roelien Bastiaanse (1993). *Studies in Aphasia*.
8. Bert Bos (1993). *Rapid User Interface Development with the Script Language Gist*.
9. Wim Kosmeijer (1993). *Barriers and Licensing*.
10. Jan-Wouter Zwart (1993). *Dutch Syntax: A Minimalist Approach*.
11. Mark Kas (1993). *Essays on Boolean Functions and Negative Polarity*.
12. Ton van der Wouden (1994). *Negative Contexts*.
13. Joop Houtman (1994). *Coordination and Constituency: A Study in Categorial Grammar*.
14. Petra Hendriks (1995). *Comparatives and Categorial Grammar*.
15. Maarten de Wind (1995). *Inversion in French*.
16. Jelly Julia de Jong (1996). *The Case of Bound Pronouns in Peripheral Romance*.
17. Sjoukje van der Wal (1996). *Negative Polarity Items and Negation: Tandem Acquisition*.
18. Anastasia Giannakidou (1997). *The Landscape of Polarity Items*.
19. Karen Lattewitz (1997). *Adjacency in Dutch and German*.
20. Edith Kaan (1997). *Processing Subject-Object Ambiguities in Dutch*.
21. Henny Klein (1997). *Adverbs of Degree in Dutch*.
22. Leonie Bosveld-de Smet (1998). *On Mass and Plural Quantification: The case of French 'des'/'du'-NPs*.
23. Rita Landeweerd (1998). *Discourse Semantics of Perspective and Temporal Structure*.
24. Mettina Veenstra (1998). *Formalizing the Minimalist Program*.
25. Roel Jonkers (1998). *Comprehension and Production of Verbs in Aphasic Speakers*.
26. Erik F. Tjong Kim Sang (1998). *Machine Learning of Phonotactics*.
27. Paulien Rijkhoek (1998). *On Degree Phrases and Result Clauses*.
28. Jan de Jong (1999). *Specific Language Impairment in Dutch: Inflectional Morphology and Argument Structure*.
29. H. Wee (1999). *Definite Focus*.
30. Eun-Hee Lee (2000). *Dynamic and Stative Information in Temporal Reasoning: Korean Tense and Aspect in Discourse*.
31. Ivilin P. Stoianov (2001). *Connectionist Lexical Processing*

32. Klarien van der Linde (2001). *Sonority Substitutions*.
33. Monique Lamers (2001). *Sentence Processing: Using Syntactic, Semantic, and Thematic Information*.
34. Shalom Zuckerman (2001). *The Acquisition of “Optional” Movement*.
35. Rob Koeling (2001). *Dialogue-Based Disambiguation: Using Dialogue Status to Improve Speech Understanding*.
36. Esther Ruigendijk (2002). *Case Assignment in Agrammatism: A Cross-Linguistic Study*.
37. Tony Mullen (2002). *An Investigation into Compositional Features and Feature Merging for Maximum Entropy-Based Parse Selection*.
38. Nanette Bienfait (2002). *Grammatica-onderwijs aan Allochtonen Jongeren*.
39. Dirk-Bart den Ouden (2002). *Phonology in Aphasia: Syllables and Segments in Level-Specific Deficits*.
40. Rienk Withaar (2002). *The Role of the Phonological Loop in Sentence Comprehension*.
41. Kim Sauter (2002). *Transfer and Access to Universal Grammar in Adult Second Language Acquisition*.
42. Laura Sabourin (2003). *Grammatical Gender and Second Language Processing: An ERP Study*.
43. Hein van Schie (2003). *Visual Semantics*.
44. Lilia Schürcks-Grozeva (2003). *Binding and Bulgarian*.
45. Stasinos Konstantopoulos (2003). *Using ILP to Learn Local Linguistic Structures*.
46. Wilbert Heeringa (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance*.
47. Wouter Jansen (2004). *Laryngeal Contrast and Phonetic Voicing: A Laboratory Phonology*.
48. Judith Rispens (2004). *Syntactic and Phonological Processing in Developmental Dyslexia*.
49. Danielle Bougaïré (2004). *L'Approche Communicative des Campagnes de Sensibilisation en Santé Publique au Burkina Faso: Les Cas de la Planification Familiale, du Sida et de l'Excision*.
50. Tanja Gaustad (2004). *Linguistic Knowledge and Word Sense Disambiguation*.
51. Susanne Schoof (2004). *An HPSG Account of Nonfinite Verbal Complements in Latin*.
52. M. Begoña Villada Moirón (2005). *Data-Driven Identification of Fixed Expressions and their Modifiability*.
53. Robbert Prins (2005). *Finite-State Pre-Processing for Natural Language Analysis*.
54. Leonoor van der Beek (2005). *Topics in Corpus-Based Dutch Syntax*.
55. Keiko Yoshioka (2005). *Linguistic and Gestural Introduction and Tracking of Referents in L1 and L2 Discourse*.
56. Sible Andringa (2005). *Form-Focused Instruction and the Development of Second Language Proficiency*.
57. Joanneke Prenger (2005). *Taal Telt! Een Onderzoek naar de Rol van Taalvaardigheid en Tekstbegrip in het Realistisch Wiskundeonderwijs*.
58. Neslihan Kansu-Yetkiner (2006). *Blood, Shame and Fear: Self-Presentation Strategies of Turkish Women's Talk about their Health and Sexuality*.
59. Mónika Z. Zempléni (2006). *Functional Imaging of the Hemispheric Contribution to Language Processing*.
60. Maartje Schreuder (2006). *Prosodic Processes in Language and Music*.
61. Hideyoshi Shiraishi (2006). *Topics in Nivkh Phonology*.
62. Tamás Biró (2006). *Finding the Right Words: Implementing Optimality Theory with Simulated Annealing*.
63. Diewuke de Goede (2006). *Verbs in Spoken Sentence Processing: Unraveling the Activation Pattern of the Matrix Verb*.
64. Eleonora Rossi (2007). *Clinic Production in Italian Agrammatism*.

65. Holger Hopp (2007). *Ultimate Attainment at the Interfaces in Second Language Acquisition: Grammar and Processing*.
66. Gerlof Bouma (2008). *Starting a Sentence in Dutch: A Corpus Study of Subject- and Object-Fronting*.
67. Julia Klitsch (2008). *Open your Eyes and Listen Carefully. Auditory and Audiovisual Speech Perception and the McGurk Effect in Dutch Speakers with and without Aphasia*.
68. Janneke ter Beek (2008). *Restructuring and Infinitival Complements in Dutch*.
69. Jori Mur (2008). *Off-line Answer Extraction for Question Answering*.
70. Lonneke van der Plas (2008). *Automatic Lexico-Semantic Acquisition for Question Answering*.
71. Arjen Versloot (2008). *Mechanisms of Language Change: Vowel Reduction in 15th Century West Frisian*.
72. Ismail Fahmi (2009). *Automatic Term and Relation Extraction for Medical Question Answering System*.
73. Tuba Yarbay Duman (2009). *Turkish Agrammatic Aphasia: Word Order, Time Reference and Case*.
74. Maria Trofimova (2009). *Case Assignment by Prepositions in Russian Aphasia*.
75. Rasmus Steinkrauss (2009). *Frequency and Function in WH Question Acquisition. A Usage-Based Case Study of German L1 Acquisition*.
76. Marjolein Deunk (2009). *Discourse Practices in Preschool. Young Children's Participation in Everyday Classroom Activities*.
77. Sake Jager (2009). *Towards ICT-Integrated Language Learning: Developing an Implementation Framework in terms of Pedagogy, Technology and Environment*.
78. Francisco Dellatorre Borges (2010). *Parse Selection with Support Vector Machines*.
79. Geoffrey Andogah (2010). *Geographically Constrained Information Retrieval*.
80. Jacqueline van Kruiningen (2010). *Onderwijsontwerp als Conversatie. Probleemoplossing in Interprofessioneel Overleg*.
81. Robert G. Shackleton (2010). *Quantitative Assessment of English-American Speech Relationships*.
82. Tim Van de Cruys (2010). *Mining for Meaning: The Extraction of Lexico-semantic Knowledge from Text*.
83. Therese Leinonen (2010). *An Acoustic Analysis of Vowel Pronunciation in Swedish Dialects*.
84. Erik-Jan Smits (2010). *Acquiring Quantification. How Children Use Semantics and Pragmatics to Constrain Meaning*.
85. Tal Caspi (2010). *A Dynamic Perspective on Second Language Development*.
86. Teodora Mehotcheva (2010). *After the Fiesta is Over: Foreign Language Attrition of Spanish in Dutch and German Erasmus Students*.
87. Xiaoyan Xu (2010). *English Language Attrition and Retention in Chinese and Dutch University Students*.
88. Jelena Prokić (2010). *Families and Resemblances*.
89. Radek Šimík (2011). *Modal Existential Wh-constructions*.
90. Katrien Colman (2011). *Behavioral and Neuroimaging Studies on Language Processing in Dutch Speakers with Parkinson's Disease*.
91. Siti Mina Tamah (2011). *A Study on Student Interaction in the Implementation of the Jigsaw Technique in Language Teaching*.
92. Aletta Kwant (2011). *Geraakt door Prentenboeken. Effecten van het Gebruik van Prentenboeken op de Sociaal-Emotionele Ontwikkeling van Kleuters*.

93. Marlies Kluck (2011). *Sentence Amalgamation*.
94. Anja Schüppert (2011). *Origin of asymmetry: Mutual Intelligibility of Spoken Danish and Swedish*.
95. Peter Nabende (2011). *Applying Dynamic Bayesian Networks in Transliteration Detection and Generation*.
96. Barbara Plank (2011). *Domain Adaptation for Parsing*.
97. Cagri Coltekin (2011). *Catching Words in a Stream of Speech: Computational Simulations of Segmenting Transcribed Child-Directed Speech*.
98. Dörte Hessler (2011). *Audiovisual Processing in Aphasic and Non-Brain-Damaged Listeners: The Whole is More than the Sum of its Parts*.
99. Herman Heringa (2012). *Appositional Constructions*.
100. Diana Dimitrova (2012). *Neural Correlates of Prosody and Information Structure*.
101. Harwintha Anjarningsih (2012). *Time Reference in Standard Indonesian Agrammatic Aphasia*.
102. Myrte Gosen (2012). *Tracing Learning in Interaction. An Analysis of Shared Reading of Picture Books at Kindergarten*.
103. Martijn Wieling (2012). *A Quantitative Approach to Social and Geographical Dialect Variation*.
104. Gisi Cannizzaro (2012). *Early Word Order and Animacy*.
105. Kostadin Cholakov (2012). *Lexical Acquisition for Computational Grammars. A Unified Model*.
106. Karin Beijering (2012). *Expressions of Epistemic Modality in Mainland Scandinavian. A Study into the Lexicalization-Grammaticalization-Pragmaticalization Interface*.
107. Veerle Baaijen (2012). *The Development of Understanding through Writing*.
108. Jacolien van Rij (2012). *Pronoun Processing: Computational, Behavioral, and Psychophysiological Studies in Children and Adults*.
109. Ankelen Schippers (2012). *Variation and Change in Germanic Long-Distance Dependencies*.
110. Hanneke Loerts (2012). *Uncommon Gender: Eyes and Brains, Native and Second Language Learners, & Grammatical Gender*.
111. Marjoleine Sloos (2013). *Frequency and Phonological Grammar: An Integrated Approach. Evidence from German, Indonesian, and Japanese*.
112. Aysa Arylova. (2013). *Possession in the Russian Clause. Towards Dynamicity in Syntax*.
113. Daniël de Kok (2013). *Reversible Stochastic Attribute-Value Grammars*.
114. Gideon Kotzé (2013). *Complementary Approaches to Tree Alignment: Combining Statistical and Rule-Based Methods*.
115. Fridah Katushemererwe (2013). *Computational Morphology and Bantu Language Learning: An Implementation for Runyakitara*.
116. Ryan C. Taylor (2013). *Tracking Referents: Markedness, World Knowledge and Pronoun Resolution*.
117. Hana Smiskova-Gustafsson (2013). *Chunks in L2 Development: A Usage-based Perspective*.
118. Milada Walková (2013). *The Aspectual Function of Particles in Phrasal Verbs*.
119. Tom O. Abuom (2013). *Verb and Word Order Deficits in Swahili-English Bilingual Agrammatic Speakers*.
120. Gülsen Yilmaz (2013). *Bilingual Language Development among the First Generation Turkish Immigrants in the Netherlands*.

121. Trevor Benjamin (2013). *Signaling Trouble: On the Linguistic Design of Other-Initiation of Repair in English Conversation*.
122. Nguyen Hong Thi Phuong (2013). *A Dynamic Usage-based Approach to Second Language Teaching*.
123. Harm Brouwer (2014). *The Electrophysiology of Language Comprehension: A Neurocomputational Model*.
124. Kendall Decker (2014). *Orthography Development for Creole Languages*.
125. Laura S. Bos (2015). *The Brain, Verbs, and the Past: Neurolinguistic Studies on Time Reference*.
126. Rimke Groenewold (2015). *Direct and Indirect Speech in Aphasia: Studies of Spoken Discourse Production and Comprehension*.
127. Huiping Chan (2015). *A Dynamic Approach to the Development of Lexicon and Syntax in a Second Language*.
128. James Griffiths (2015). *On Appositives*.
129. Pavel Rudnev (2015). *Dependency and Discourse-Configurationality: A Study of Avar*.
130. Kirsten Kolstrup (2015). *Opportunities to Speak. A Qualitative Study of a Second Language in Use*.
131. Gülistan Güneş (2015). *Deriving Prosodic Structures*.
132. Cornelia Lahmann (2015). *Beyond Barriers. Complexity, Accuracy, and Fluency in Long-Term L2 Speakers' Speech*.
133. Sri Wachyunni (2015). *Scaffolding and Cooperative Learning: Effects on Reading Comprehension and Vocabulary Knowledge in English as a Foreign Language*.
134. Albert Walsweer (2015). *Ruimte voor Leren. Een Etnografisch Onderzoek naar het Verloop van een Interventie Gericht op Versterking van het Taalgebruik in een Knowledge Building Environment op Kleine Friese Basisscholen*.
135. Aleyda Lizeth Linares Calix (2015). *Raising Metacognitive Genre Awareness in L2 Academic Readers and Writers*.
136. Fathima Mufeeda Irshad (2015). *Second Language Development through the Lens of a Dynamic Usage- Based Approach*.
137. Oscar Strik (2015). *Modelling Analogical Change. A History of Swedish and Frisian Verb Inflection*.
138. He Sun (2015). *Predictors and Stages of Very Young Child EFL Learners' English Development in China*.
139. Marieke Haan (2015). *Mode Matters. Effects of Survey Modes on Participation and Answering Behavior*.
140. Nienke Houtzager (2015). *Bilingual Advantages in Middle-Aged and Elderly Populations*.
141. Noortje Joost Venhuizen (2015). *Projection in Discourse: A Data-Driven Formal Semantic Analysis*.
142. Valerio Basile (2015). *From Logic to Language: Natural Language Generation from Logical Forms*.
143. Jinxing Yue (2016). *Tone-Word Recognition in Mandarin Chinese: Influences of Lexical-Level Representations*.
144. Seçkin Arslan (2016). *Neurolinguistic and Psycholinguistic Investigations on Evidentiality in Turkish*.
145. Rui Qin (2016). *Neurophysiological Studies of Reading Fluency. Towards Visual and Auditory Markers of Developmental Dyslexia*.
146. Kashmire Stec (2016). *Visible Quotation: The Multimodal Expression of Viewpoint*.
147. Yinxing Jin (2016). *Foreign Language Classroom Anxiety: A Study of Chinese University Students of Japanese and English over Time*.
148. Joost Hurkmans (2016). *The Treatment of Apraxia of Speech. Speech and Music Therapy, an Innovative Joint Effort*.

149. Franziska Köder (2016). *Between Direct and Indirect Speech: The Acquisition of Pronouns in Reported Speech*.
150. Femke Swarte (2016). *Predicting the Mutual Intelligibility of Germanic Languages from Linguistic and Extra-Linguistic Factors*.
151. Sanne Kuijper (2016). *Communication Abilities of Children with ASD and ADHD. Production, Comprehension, and Cognitive Mechanisms*.
152. Jelena Golubović (2016). *Mutual Intelligibility in the Slavic Language Area*.
153. Nynke van der Schaaf (2016). "Kijk Eens Wat Ik Kan!" *Sociale Praktijken in de Interactie Tussen Kinderen van 4 tot 8 Jaar in de Buitenschoolse Opvang*.
154. Simon Šuster (2016). *Empirical Studies on Word Representations*.
155. Kilian Evang (2016). *Cross-Lingual Semantic Parsing with Categorial Grammars*.
156. Miren Arantzeta Pérez (2017). *Sentence Comprehension in Monolingual and Bilingual Aphasia: Evidence from Behavioral and Eye-Tracking Methods*.
157. Sana-e-Zehra Haidry (2017). *Assessment of Dyslexia in the Urdu Language*.
158. Srđan Popov (2017). *Auditory and Visual ERP Correlates of Gender Agreement Processing in Dutch and Italian*.
159. Molood Sadat Safavi (2017). *The Competition of Memory and Expectation in Resolving Long-Distance Dependencies: Psycholinguistic Evidence from Persian Complex Predicates*.
160. Christopher Bergmann (2017). *Facets of Native-Likeness: First-Language Attrition among German Emigrants to Anglophone North America*.
161. Stefanie Keulen (2017). *Foreign Accent Syndrome: A Neurolinguistic Analysis*.
162. Franz Manni (2017). *Linguistic Probes into Human History*.
163. Margreet Vogelzang (2017). *Reference and Cognition: Experimental and Computational Cognitive Modeling Studies on Reference Processing in Dutch and Italian*.
164. Johannes Bjerva (2017). *One Model to Rule them All. Multitask and Multilingual Modelling for Lexical Analysis*.
165. Dieke Oele (2018). *Automated Translation with Interlingual Word Representations*.
166. Lucas Seuren (2018). *The Interactional Accomplishment of Action*.
167. Elisabeth Borleffs (2018). *Cracking the Code — Towards Understanding, Diagnosing and Remediating Dyslexia in Standard Indonesian*.
168. Mirjam Günther-van der Meij (2018). *The Impact of Degree of Bilingualism on L3 Development: English Language Development in Early and Later Bilinguals in the Frisian Context*.
169. Ruth Koops van 't Jagt (2018). *Show, Don't just Tell: Photo Stories to Support People with Limited Health Literacy*.
170. Bernat Bardagil-Mas (2018). *Case and Agreement in Panará*.
171. Jessica Overweg (2018). *Taking an Alternative Perspective on Language in Autism*.
172. Lennie Donné (2018). *Convincing through Conversation: Unraveling the Role of Interpersonal Health Communication in Health Campaign Effectiveness*.

173. Toivo Glatz (2018). *Serious Games as a Level Playing Field for Early Literacy: A Behavioural and Neurophysiological Evaluation*.
174. Ellie van Setten (2019). *Neurolinguistic Profiles of Advanced Readers with Developmental Dyslexia*.
175. Anna Pot (2019). *Aging in Multilingual Netherlands: Effects on Cognition, Wellbeing and Health*.
176. Audrey Rousse-Malpat (2019). *Effectiveness of Explicit vs. Implicit L2 Instruction: A Longitudinal Classroom Study on Oral and Written Skills*.
177. Rob van der Goot (2019). *Normalization and Parsing Algorithms for Uncertain Input*.
178. Azadeh Elmianvari (2019). *Multilingualism, Facebook and the Iranian Diaspora*.
179. Joëlle Ooms (2019). “*Don’t Make my Mistake*”: Narrative Fear Appeals in Health Communication.
180. Annerose Willemse (2019). *The Floor is Yours: A Conversation Analytic Study of Teachers’ Conduct Facilitating Whole-Class Discussions around Texts*.
181. Frans Hiddink (2019). *Early Childhood Problem-Solving Interaction: Young Children’s Discourse During Small-Group Work in Primary School*.
182. Hessel Haagsma (2020). *A Bigger Fish to Fry: Scaling up the Automatic Understanding of Idiomatic Expressions*.
183. Juliana Andrade Feiden (2020). *The Influence of Conceptual Number in Coreference Establishing: An ERP Study on Brazilian and European Portuguese*.
184. Sirkku Lesonen (2020). *Valuing Variability: Dynamic Usage-Based Principles in the L2 Development of Four Finnish Language Learners*.
185. Nathaniel Larney (2020). *A Neurolinguistic Approach to the Processing of Resumption in Akan Focus Constructions*.
186. Bernard Amadeus Jaya Jap (2020). *Syntactic Frequency and Sentence Processing in Standard Indonesian*.
187. Ting Huang (2020). *Learning an L2 and L3 at the Same Time: Help or Hinder?*
188. Anke Herder (2020). *Peer Talk in Collaborative Writing of Primary School Students: A Conversation Analytic Study of Student Interaction in the Context of Inquiry Learning*.
189. Ellen Schep (2020). *Attachment in Interaction: A Conversation Analytic Study on Dinner Conversations with Adolescents in Family-Style Group Care*.
190. Yulia Akinina (2020). *Individual Behavioural Patterns and Neural Underpinnings of Verb Processing in Aphasia*.
191. Camila Martinez Rebolledo (2020). *Comprehending the Development of Reading Difficulties in Children with SLI*.
192. Jakolien den Hollander (2021). *Distinguishing a Phonological Encoding Disorder from Apraxia of Speech in Individuals with Aphasia by Using EEG*.
193. Rik van Noord (2021). *Character-Based Neural Semantic Parsing*.
194. Anna de Koster (2021). *Acting Individually or Together? An Investigation of Children’s Development of Distributivity*.
195. Frank Tsiwah (2021). *Time, Tone and the Brain: Behavioral and Neurophysiological Studies on Time Reference and Grammatical Tone in Akan*.
196. Amélie la Roi (2021). *Idioms in the Aging Brain*.

197. Nienke Wolthuis (2021). *Language Impairments and Resting-State EEG in Brain Tumour Patients: Revealing Connections*.
198. Nienke Smit (2021). *Get It Together: Exploring the Dynamics of Teacher-Student Interaction in English as a Foreign Language Lessons*.
199. Svetlana Averina (2021). *Bilateral Neural Correlates of Treatment-Induced Changes in Chronic Aphasia*.
200. Wilasinee Siriboonpipattana (2021). *Neurolinguistic Studies on the Linguistic Expression of Time Reference in Thai*.
201. Irene Graafsma (2021). *Computer Programming Skills: A Cognitive Perspective*.
202. Pouran Seifi (2021). *Processing and Comprehension of L2 English Relative Clauses by Farsi Speakers*.
203. Hongying Peng (2021). *A Holistic Person-Centred Approach to Mobile-Assisted Language Learning*.
204. Nermina Cordalija (2021). *Neurolinguistic and Psycholinguistic Approaches to Studying Tense, Aspect, and Unaccusativity*.
205. Aida Salčić (2021). *Agreement Processing in Dutch Adults with Dyslexia*.
206. Eabele Tjepkema (2021). *Exploring Content-Based Language Teaching Practices to Stimulate Language Use in Grades 7 and 8 of Frisian Trilingual Primary Education*.
207. Liefke Reitsma (2021). *Bilingualism and Contact-Induced Language Change: Exploring Variation in the Frisian Verbal Complex*.
208. Steven Gilbers (2021). *Ambitionz az a Ridah: 2Pac's Changing Accent and Flow in Light of Regional Variation in African-American English Speech and Hip-Hop Music*.
209. Leanne Nagels (2021). *From Voice to Speech: The Perception of Voice Characteristics and Speech in Children with Cochlear Implants*.
210. Vasilisa Verkhodanova (2021). *More than Words: Recognizing Speech of People with Parkinson's Disease*.
211. Liset Rouweler (2021). *The Impact of Dyslexia in Higher Education*.
212. Maaike Pulles (2021). *Dialogic Reading Practices: A Conversation Analytic Study of Peer Talk in Collaborative Reading Activities in Primary School Inquiry Learning*.
213. Agnes M. Engbersen (2022). *Assisting Independent Seniors with Morning Care: How Care Workers and Seniors Negotiate Physical Cooperation through Multimodal Interaction*.
214. Ryssa Moffat (2022). *Recognition and Cortical Haemodynamics of Vocal Emotions — An fNIRS Perspective*.
215. Diane Mézière (2022). *Using Eye Movements to Develop an Ecologically-Valid AI Measure of Reading Comprehension*.
216. Ann-Katrin Ohlerth (2022). *Improving Preoperative nTMS with a Dual-Task Protocol: The Contribution of Action Naming to Language Mapping*.
217. Masha Medvedeva (2022). *Identification, Categorisation and Forecasting of Court Decisions*.
218. Jidde Jacobi (2022). *Coordination and Timing of Speech Gestures in Parkinson's Disease*.
219. Dorothée Hoppe (2022). *Linear Order in Language: An Error-Driven Learning Account*.
220. Janine Strandberg (2022). *Phonetic and Lexical Variation in Finland-Swedish*.
221. Wim Gombert (2022). *From "Learning to Use" Towards "Using to Learn"? Long-Term Effects of Structure-Based versus Dynamic Usage-Based Programs for French*.

222. Marlon Ruwette (2023). *Reading Promotion of Fiction and Nonfiction Texts: Determinants of Elementary School Teachers' Reading Promotion Behavior and their Students' Leisure Time Reading*.
223. Ahmet Üstün (2023). *Multilingual Learning and Adaptation for Neural Language Models*.
224. Suzan Dilara Tokaç (2023). *An Investigation of Heritage Language Speakers of Turkish: Evidentiality Processing and the Effects of Language Experience*.
225. Marita K. Everhardt (2023). *Non-Native Prosody Perception in Cochlear Implant-Simulated Speech*.
226. Jeroen Steenbakkers (2023). *Speelse Toewijding. Een Pedagogisch-Didactisch Onderzoek naar Schrijfstijl- en Formuleeronderwijs in Klas 3 en 4 Havo/Vwo*.
227. Nienke Meulman (2023). *Second Language Processing: Electrophysiological Studies and Data Analysis Methods*.
228. Antje Sien Lisanne Knol (2023). *Doing Therapies: On the Tangibility of Theory in Interactional Practices*.
229. Sofia Bimpikou (2023). *Inside Characters' Minds: The Role of Reports in Narrative Perspective Taking*.
230. Pauline Cuperus (2023). *Aphasia Therapy Software: Research, Development, and Implementation*.
231. Elizabeth Heredia Murillo (2023). *The Semantics and the L1 Acquisition of Spanish Scalar (Focus) Particles: The Case of (ni) *siquiera* and *incluso**.
232. Mara van der Ploeg (2023). *Language Learning Never Gets Old: Learning a New Language in Later Life*.
233. Martijn Bartelds (2023). *Representing Low-Resource Languages and Dialects: Improved Neural Methods for Spoken Language Processing*.
234. Yingjie Wang (2023). *Unpacking the Label "Made in China": A Critical Discourse Analysis of Chinese and US News Discourse*.
235. Prajit Dhar (2024). *Improving Neural Machine Translation of Languages with Little Data and Rich Morphology*.
236. Effy Ntemou (2024). *Effects of Linguistic Properties in Preoperative Language Mapping: Insights from VLSM and Navigated TMS*.
237. Huiyuan Lai (2024). *Neural Text Rewriting: Style Transfer, Figurative Language, and Beyond*.
238. Chunliu Wang (2024). *Multilingual Semantic Parsing and Generation with Neural Models*.
239. Lukas Edman (2024). *The Little Data that Could: Making the Most of Low-Resource Natural Language Processing*.
240. Jelle Brouwer (2024). *Learning for Life: Investigating the Potential of Language Learning as a Non Pharmacological Intervention in Older Adults with (Past) Depression*.
241. Daan van Soeren (2024). *Chronicle of a Linguistic Form Foretold: How Human Efficiency and Prosodic Factors Explain the Distribution of Labials and Coronals in Spanish, English, and Dutch*.
242. Wietse de Vries (2024). *Evaluation and Adaptation of Neural Language Models for Under-Resourced Languages*.
243. Suzanne Dekker (2024). *Multilingualism in Heads, Hearts, and Hands: Attitudes Towards the Development, Implementation, and Evaluation of Multilingual Approaches in Frisian Primary Schools*.
244. Cheyenne Svaldi (2024). *An Investigation of Language Learning and Processing in Children with Developmental Language Disorder and Pediatric Cerebellar Tumor Survivors*.
245. Gosse Minnema (2024). *Perspective Matters: Framing Events in Language and Society*.

246. Solveig Castelli (2024). *The Planning and Preparation of Words for Language Production*.
247. Luc Koning (2024). *Redzaamheidslezen*.
248. Lourens Visser (2024). *Adverbs of Degree in Early West Germanic*.
249. Floor van den Berg (2024). *Introducing Bilingual Experience as a Healthy Cognitive Aging Tool*.
250. Sara Shoghi Javan (2024). *The Impact of Visual Contextual Information and Verb Argument Structure on Syntactic Parsing Preferences: A Behavioral-Electrophysiological Investigation of Ambiguous Structures in Dutch*.
251. Andreas Hiemstra (2025). *Morphosyntactic Transfer from German and English into Dutch in Third Language Acquisition: A Study on Dutch Superlatives and Two-Verb Clusters*.
252. Joke van Balen (2025). *Making Room for Subjectification during Classroom Discussions. A Conversation Analytic Study of Interactional Practices in the Dutch Language Classroom*.
253. Raoul Buurke (2025). *Frisian and Low Saxon in Flux*.
254. Teja Rebernik (2024). *Acoustic and Kinematic Insights into Articulatory Sensorimotor Control in Parkinson's Disease*.
255. Roelant Ossewaarde (2025). *Automated Measurements of Fluency, Syntax and Semantics in the Language of Persons with Primary Progressive Aphasia*.
256. Joanna Porkert (2025). *Can She be a Truckee? Pronoun and Gender Processing in the Multilingual Brain*.
257. Nelleke Jansen (2025). *In Tune: Exploring the Role of Musical Abilities in Second-Language Prosodic Processing and Production*.
258. Matt Drury (2025). *Local to Global Ecolinguistics: In the Name of Sustainability*.
259. May Yue Wu (2025). *Mind the Metrics: CAF Measure Reliability and the Implications for L2 Studies*.
260. Miguel Angel Santin Schulz (2025). *Scalar Change and Event Endpoints in Language and Cognition*.
261. Marjolein Wietske Talsma (2025). *Argument Structure as a Function of Merge: From UTAH to IDAHO*.
262. Demian Stoianov (2025). *Reading with Emojis: A Behavioural and Neurophysiological Investigation*.
263. Gabriele Sarti (2025). *From Insights to Impact: Actionable Interpretability for Neural Machine Translation*.

GRODIL

Center for Language and Cognition Groningen (CLCG)  
 P.O. Box 716  
 9700 AS Groningen  
 The Netherlands