

T.C.
İSTANBUL 29 MAYIS ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
MÜTERCİM VE TERCÜMANLIK ANABİLİM DALI

**FROM PROMPTS TO PRECISION: A COMPARATIVE
EXPERIMENTAL STUDY OF TRANSLATION QUALITY
OUTPUTS WITH FOCUS ON ROLE-BASED (PERSONA)
PROMPTING FOR MEDICAL TEXTS**

(YÜKSEK LİSANS TEZİ)

Mehmet Ferit DOĞAN

Danışman:

Prof. Dr. Işın ÖNER

İSTANBUL

2025

T. C.
İSTANBUL 29 MAYIS ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
MÜTERCİM VE TERCÜMANLIK ANABİLİM DALI

**FROM PROMPTS TO PRECISION:
A COMPARATIVE EXPERIMENTAL STUDY OF
TRANSLATION QUALITY OUTPUTS
WITH FOCUS ON ROLE-BASED (PERSONA)
PROMPTING FOR MEDICAL TEXTS**

(YÜKSEK LİSANS TEZİ)

Mehmet Ferit DOĞAN

Danışman:
Prof. Dr. Işın ÖNER

İSTANBUL

2025

TEZ ONAY SAYFASI

T. C.

İSTANBUL 29 MAYIS ÜNİVERSİTESİ

SOSYAL BİLİMLER ENSTİTÜSÜ MÜDÜRLÜĞÜNE

Mütercim ve Tercümanlık Anabilim Dalı, Çeviribilim (İngilizce) Tezli Yüksek Lisans Programı'nda 200522YL03 numaralı Mehmet Ferit DOĞAN'ın hazırladığı “*From Prompts to Precision: A Comparative Experimental Study of Translation Quality Output with Focus on Role-based (Persona) Prompting for Medical Texts*” konulu yüksek lisans tezi ile ilgili tez savunma sınavı, 03/06/2025 günü 11:30 – 12:15 saatleri arasında yapılmış, sorulan sorulara alınan cevaplar sonunda adayın tezinin başarılı olduğuna oy birliği ile karar verilmiştir.

Prof. Dr. Işın ÖNER
İstanbul 29 Mayıs Üniversitesi
(Tez Danışmanı ve Sınav Komisyonu Başkanı)

Prof. Dr. Ayşe Banu KARADAĞ
Yıldız Teknik Üniversitesi

Doç. Dr. Nilüfer ALİMEN
Yıldız Teknik Üniversitesi

BEYAN

Bu tezin yazılmasında bilimsel ahlak kurallarına uyulduğunu, başkalarının eserlerinden yararlanılması durumunda bilimsel normlara uygun olarak atıfta bulunulduğunu, kullanılan verilerde herhangi bir tahrifat yapılmadığını, tezin herhangi bir kısmının bu üniversite veya başka bir üniversitedeki başka bir tez çalışması olarak sunulmadığını beyan ederim.

Mehmet Ferit DOĞAN

03/06/2025

ABSTRACT

With the advancements in artificial intelligence-driven machine translation activities, designing prompts for large language models for translation task is an area that recently has gained interest in translation literature. The effectiveness of assigning personas into prompts, however, is a subject that needs further research in medical/pharmaceutical texts domain. In response to this gap, this study aims to evaluate the impact of offering role-based (persona) prompts to large language models. By doing so, the research assesses the quality of translation outputs of role-prompted large language models against 3 different source texts with pharmaceutical content, while comparing these outputs to zero-shot prompted models alongside the outputs from long-rooted conventional neural machine translation system: Google Translate. While evaluating automated metric scores such as BLEU (Bilingual Evaluation Understudy) and COMET (Crosslingual Optimized Metric for Evaluation of Translation) from the machine translation outputs, I performed TQE (Translation Quality Evaluation) analysis through a customized TQE error-typology tool, motivated and inspired by MQM (Multidimensional Quality Metrics) framework. Therefore, this study tries to discover how effective role-based prompts are in medical/pharmaceutical domain translation with both a quantitative and also a qualitative approach, as I believe that such a hybrid approach is essential for a solid oversight of quality. Lastly, the study will include discussions over the widely-discussed concepts of machine translation ethics, translator competence, and human-machine collaboration.

Keywords:

Role-based Prompting, Large Language Models, Machine Translation, Translation Quality Evaluation, Medical Texts, Machine Ethics

ÖZ

Yapay zekâ tabanlı makine çevirisi etkinliğindeki gelişmelerle birlikte, çeviri eyleminde büyük dil modelleri için istem (prompt) tasarlama yakın dönemde çeviri alanyazınında ilgi görmüş bir alandır. Ne var ki söz konusu istemlere rol (persona) atamanın yaratacağı etki, tıp/ilaç metinleri bağlamında detaylı araştırılma ihtiyacı taşımaktadır. Alanyazındaki bu boşluğa yanıt olarak, bu çalışmada büyük dil modellerine rol temelli istemler sunmanın yaratacağı etki incelenmektedir. Söz konusu incelemeyi yaparken bu çalışma; ilaç metinleri bağlamındaki üç farklı kaynak metin üzerinden, rol temelli istem sunulan büyük dil modellerinin çeviri kalite çıktılarını sıfır atışlı (zero-shot) istem sunulan modellerin çeviri çıktılarıyla kıyaslamaktadır. Ayrıca bu kıyaslamaya, uzun süredir çeviri etkinliğinde yer edinmiş olan, konvansiyonel bir nöral makine çevirisi aracı olan Google Translate’in kalite çıktıları da dahil edilmiştir. Çalışmadaki makine çevirisi çıktıları; BLEU (Bilingual Evaluation Understudy) ve COMET (Crosslingual Optimized Metric for Evaluation of Translation) gibi otomatik kalite ölçütleri aracılığıyla kalite puanlamasına tabi olmuştur. Bu ölçütlerin yanı sıra, MQM’den (Multidimensional Quality Metrics) esinlenilerek tarafımda oluşturulan uyarlanılmış bir çeviri hata tipolojisi aracı kullanılarak çeviri kalite ölçümü gerçekleştirilmiştir. Böylece bu çalışma, tıp/ilaç metinleri bağlamında rol temelli istemlerin çeviri eyleminde ne kadar etkili olduğunu hem nicel hem de nitel bir yaklaşımla keşfetmeyi hedeflemektedir. Zira güçlü bir kalite gözetimi için bu ikili yaklaşımın gerektiğine inanmaktayım. Son olarak, bu çalışmanın tartışma bölümünde makine çevirisi etiği, çevirmen yetkinliği, insan-makine iş birliği gibi yaygın tartışılan kavramlara yer verilmiştir.

Anahtar Kelimeler:

Rol Temelli İstem, Büyük Dil Modelleri, Makine Çevirisi, Çeviride Kalite
Değerlendirmesi, Tıp Metinleri, Makine Etiği

ACKNOWLEDGMENTS

I would first like to express my profound gratitude to my thesis advisor Prof. Dr. Işın ÖNER for having always been a massive source of guidance to me, for our fruitful discussions, and for her inspiring perspectives. Being her student is an enormous luck, and she always helped me to be motivated for life-long learning on numerous topics with her invaluable teachings.

I would also like to offer heartfelt thanks to Prof. Dr. Ayşe Banu KARADAĞ for her invaluable contributions and teachings to me and my peers. She always shed light on how to be prepared to become effective and more improved in academic writing, reading, and speeches.

I would like to thank to Assoc. Prof. Dr. Nilüfer ALİMEN for her guidance on research methodologies and for all her contributions in academic reading and writing.

I want to thank my dear friends Sena EKİNCİ, Özümsu ALAGÖZ and Volkan DEDE for sharing their experiences from their academic journeys, for their constant support and joyful conversations, and motivating me.

I want to extend my special and heartfelt thanks my dear mother Sibel DOĞAN and my father Hamdi Ferit DOĞAN for spreading positivity and contagious laughs alongside making me feel appreciated and always supported throughout my academic journey.

Last but not the least, I am beyond thankful to my dear brother Muhammet Süreyya DOĞAN, the existence of whom makes me smile, shine and be happy in life.

TABLE OF CONTENTS

TEZ ONAY SAYFASI.....	ii
BEYAN.....	iii
ABSTRACT	iv
ÖZ	v
ACKNOWLEDGMENTS	vi
TABLE OF CONTENTS	vii
LIST OF ABBREVIATIONS.....	xi
LIST OF TABLES.....	xiii
LIST OF FIGURES	xv
1. INTRODUCTION	1
1.1 Research Questions.....	5
2. AIM OF THE RESEARCH.....	6
3. DEFINING THE RESEARCH AREA	8
3.1 Prompting LLMs	8
3.1.1 Zero-shot Prompting.....	9
3.1.2 Role-based/Persona Prompting.....	9
3.2 Medical Texts as a Domain	10
3.3 Importance of Medical Translation	12
4. LITERATURE REVIEW	14
4.1 Machine Translation (MT) Optimization via Prompting.....	14
4.2 Persona Prompting in Large Language Models (LLMs).....	15
4.3 Generative AI and Machine Translation in Medical Domain.....	16

5. METHODOLOGY	20
5.1 First Pillar: Pre-production Process of Translated Product	20
5.1.1 ST-1	21
5.1.1.1 Reference Text-1.....	23
5.1.1.2 Zero-shot (Raw) Prompt	24
5.1.1.3 Role-based Prompt.....	24
5.1.2 ST-2	25
5.1.2.1 Reference Text-2.....	26
5.1.2.2 Zero-shot (Raw) Prompt	27
5.1.2.3 Role-based Prompt.....	27
5.1.3. ST-3	28
5.1.3.1 Reference Text-3.....	30
5.1.3.2 Zero-shot (Raw) Prompt	30
5.1.3.3 Role-based Prompt.....	31
5.1.4 Actuators Deployed	31
5.1.5 Qualitative Framework: TQE (Translation Quality Evaluation) Penalty Score Tool .	32
5.1.5.1 TQE Spreadsheet	32
5.1.5.2 Error-typology Legend	35
5.1.5.3 Scoring Matrix	35
5.1.5.4 Instructions for Use for TQE Tool.....	36
5.2 Second Pillar: Production Process of Translated Product	37
5.2.1 Quantitative Approach: Automatic MT Evaluation Metrics for ST-1’s Translation Outputs.....	37
5.2.2 Qualitative Approach: Customized Error-typology Framework (ST-1).....	40
5.2.3 Quantitative Approach: Automatic MT Evaluation Metrics for ST-2’s Translation Outputs.....	41
5.2.4 Qualitative Approach: Customized Error-typology Framework (ST-2).....	42
5.2.5 Quantitative Approach: Automatic MT Evaluation Metrics for ST-3’s Translation Outputs.....	43

5.2.6 Qualitative Approach: Customized Error-typology Framework (ST-3).....	44
5.3 Commentary on Findings	44
5.3.1.1 Google Translate – raw	45
5.3.1.2 GPT-4 – zero-shot.....	48
5.3.1.3 GPT-4 – RBP	51
5.3.1.5 Gemini-2.0-flash – RBP	55
5.3.1.6 DeepSeek – R1 – zero-shot.....	57
5.3.1.8 Claude-3.5-Sonnet – zero-shot	62
5.3.1.9 Claude-3.5-Sonnet– RBP.....	64
5.3.2 Overall Results Overview for ST-1’s TT Outputs.....	66
5.3.3 Analysis on ST-2’s TT Outputs	67
5.3.3.1 Google Translate – raw	67
5.3.3.2 GPT-4 – zero-shot.....	70
5.3.3.3 GPT-4 – RBP	72
5.3.3.4 Gemini-2.0-flash – zero-shot	74
5.3.3.5 Gemini-2.0-flash – RBP	76
5.3.3.6 DeepSeek – R1 – zero-shot.....	78
5.3.3.7 DeepSeek – R1 – RBP	80
5.3.3.8 Claude-3.5-Sonnet – zero-shot	82
5.3.3.9 Claude-3.5-Sonnet – RBP.....	83
5.3.4 Overall Results Overview for ST-2’s TT Outputs.....	86
5.3.5 Analysis on ST-3’s TT Outputs	86
5.3.5.1 Google Translate – raw	86
5.3.5.2 GPT-4 – zero-shot.....	90
5.3.5.3 GPT-4 – RBP	94
5.3.5.4 Gemini-2.0-flash – zero-shot	97
5.3.5.5 Gemini-2.0-flash – RBP	99
5.3.5.6 DeepSeek - R1 – zero-shot	102
5.3.5.7 DeepSeek - R1 – RBP.....	104

5.3.5.8 Claude-3.5-Sonnet – zero-shot	107
5.3.5.9 Claude-3.5-Sonnet – RBP	109
5.3.6 Overall Results Overview for ST-3’s TT Outputs.....	112
6. DISCUSSION	113
6.1 Text-type and Function Analysis.....	113
6.2 Domain-specific Oversight.....	115
6.3 Would Human – Machine Collaboration be Possible?.....	116
6.4 Qualifications Expected from a Medical Translator.....	119
6.5 Ethical Considerations & Risks Posed by LLMs	121
7. CONCLUDING REMARKS	125
REFERENCES	128
APPENDIX A – EXTRACT FROM ST-1 (PAGE 12).....	135
APPENDIX B – EXTRACT FROM REFERENCE TEXT-1 (PAGE 30).....	136
APPENDIX C – EXTRACT FROM ST-2 (PAGE 4).....	137
APPENDIX D – EXTRACT FROM REFERENCE TEXT-2 (PAGE 4)	138
APPENDIX E – EXTRACT FROM ST-3 (PAGES 6-7).....	139
APPENDIX F – EXTRACT FROM REFERENCE TEXT-3 (PAGE 6)	141
ÖZGEÇMİŞ	142

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy
COMET	Crosslingual Optimized Metric for Evaluation of Translation
DNN	Deep Neural Networks
EMA	European Medicines Agency
FDA	Food and Drug Administration
GPT	Generative Pre-trained Transformer
HOPE	Human-Oriented and Professional Evaluation
ISO	International Standard Organization
LGP	Language for General Purposes
LLM	Large Language Model
LSP (1)	Language for Specific Purposes
LSP (2)	Language Service Provider
METEOR	Metric for Evaluation of Translation with Explicit Ordering
ML	Machine Learning
MQM	Multidimensional Quality Metrics
MT	Machine Translation
MTPE	Machine Translation Post-editing
NLP	Natural Language Processing
NMT	Neural Machine Translation
PIL	Patient Information Leaflet
Q&A	Question and Answer
RBP	Role-based Prompt
RNN	Recurrent Neural Networks

ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SmPC	Summary of Product Characteristics
SMT	Statistical Machine Translation
ST	Source Text
TSP	Translation Service Provider
TT	Target Text



LIST OF TABLES

Table 1: ST-1	21
Table 2: Reference Text for ST-1	23
Table 3: Zero-shot Prompt Provided to LLMs for Translation of ST-1	24
Table 4: Role-based Prompt Provided to LLMs for Translation of ST-1	24
Table 5: ST-2	25
Table 6: Pre-edited Reference Text for ST-2.....	27
Table 7: Zero-shot Prompt Provided to LLMs for Translation of ST-2	27
Table 8: Role-based Prompt Provided to LLMs for Translation of ST-2.....	28
Table 9: ST-3	29
Table 10: Reference Text for ST-3	30
Table 11: Zero-shot Prompt Provided to LLMs for Translation of ST-3	31
Table 12: Role-based Prompt Provided to LLMs for Translation of ST-3.....	31
Table 13: Metric Evaluation Results of ST-1’s Candidate Texts	37
Table 14: Visualization of 1-gram Matching in BLEU Metric Calculation	39
Table 15: TQE Overview of Penalty Scores of ST-1	40
Table 16: Metric Evaluation Results of ST-2’s Candidate Texts	41
Table 17: TQE Overview of Penalty Scores of ST-2	42
Table 18: Metric Evaluation Results of ST-3’s Candidate Texts	43
Table 19: TQE Overview of Penalty Scores of ST-3	44
Table 20: ST-1, Reference Text for ST-1, and MT Output (Google Translate - raw).....	45
Table 21: ST-1, Reference text for ST-1, and MT Output (GPT-4 – zero-shot)	48
Table 22: ST-1, Reference Text for ST-1, and MT Output (GPT-4 - RBP).....	51
Table 23: ST-1, Reference Text for ST-1, and MT Output (Gemini-2.0-Flash – zero-shot)	53
Table 24: ST-1, Reference Text for ST-1, and MT Output (Gemini-2.0-flash - RBP)	55
Table 25: ST-1, Reference Text for ST-1, and MT Output (DeepSeek-R1 – zero-shot)	57
Table 26: ST-1, Reference Text for ST-1, and MT Output (DeepSeek-R1 - RBP)	60

Table 27: ST-1, Reference Text for ST-1, and MT Output (Claude-3.5-Sonnet – zero-shot)	62
Table 28: ST-1, Reference Text for ST-1, and MT Output (Claude-3.5-Sonnet - RBP).....	64
Table 29: ST-2, Reference Text for ST-2, and MT Output (Google Translate - raw).....	67
Table 30: ST-2, Reference Text for ST-2, and MT Output (GPT-4 – zero-shot).....	70
Table 31: ST-2, Reference Text for ST-2, and MT Output (GPT-4 - RBP).....	72
Table 32: ST-2, Reference Text for ST-2, and MT Output (Gemini-2.0-flash – zero-shot)	74
Table 33: ST-2, Reference Text for ST-2, and MT Output (Gemini-2.0-flash - RBP)	76
Table 34: ST-2, Reference Text for ST-2, and MT Output (DeepSeek-R1 – zero-shot)	78
Table 35: ST-2, Reference Text for ST-2, and MT Output (DeepSeek-R1 - RBP)	80
Table 36: ST-2, Reference Text for ST-2, and MT Output (Claude-3.5-Sonnet – zero-shot)	82
Table 37: ST-2, Reference Text for ST-2, and MT Output (Claude-3.5-Sonnet - RBP).....	83
Table 38: ST-3, Reference Text for ST-3, and MT Output (Google Translate - raw).....	86
Table 39: ST-3, Reference Text for ST-3, and MT Output (GPT-4 – zero-shot).....	90
Table 40: ST-3, Reference Text for ST-3, and MT Output (GPT-4 - RBP).....	94
Table 41: ST-3, Reference Text for ST-3, and MT Output (Gemini-2.0-flash – zero-shot)	97
Table 42: ST-3, Reference Text for ST-3, and MT Output (Gemini-2.0-flash – RBP).....	99
Table 43: ST-3, Reference Text for ST-3, and MT Output (DeepSeek – R1 – zero-shot)	102
Table 44: ST-3, Reference Text for ST-3, and MT Output (DeepSeek – R1 – RBP)	104
Table 45: ST-3, Reference Text for ST-3, and MT Output (Claude-3.5-Sonnet – zero-shot)	107
Table 46: ST-3, Reference Text for ST-3, and MT Output (Claude-3.5-Sonnet – RBP)..	109

LIST OF FIGURES

Figure 1: Screenshot of TQE Spreadsheet Template.....	33
Figure 2: TQE Penalty Result of an Example-based Scenario	34
Figure 3: Screenshot of TQE Spreadsheet Legend.....	35
Figure 4: Screenshot of Scoring Matrix.....	36
Figure 5: Screenshot of Instructions	36
Figure 6: Interface of BLEU Evaluator of Tilde MT with Uploaded Candidate and Reference Texts	39
Figure 7: Error-typology Chart (Google Translate - raw)	46
Figure 8: Error-typology Chart (GPT-4 – zero-shot)	50
Figure 9: Error-typology Chart (GPT-4 - RBP).....	52
Figure 10: Error-typology Chart (Gemini-2.0-flash - zero-shot).....	54
Figure 11: Error-typology Chart (Gemini-2.0-flash - RBP).....	56
Figure 12: Error-typology chart (DeepSeek-R1 – zero-shot).....	58
Figure 13: Error-typology Chart (DeepSeek-R1 - RBP)	61
Figure 14: Error-typology Chart (Claude-3.5-Sonnet – zero-shot)	63
Figure 15: Error-typology Chart (Claude-3.5-Sonnet - RBP)	65
Figure 16: Heatmap Overview of Quantitative and Qualitative Metric Results of ST-1’s TT Outputs.....	66
Figure 17: Error-typology Chart (Google Translate - raw)	68
Figure 18: Error-typology Chart (GPT-4 – zero-shot)	71
Figure 19: Error-typology Chart (GPT-4 - RBP)	74
Figure 20: Error-typology Chart (Gemini-2.0-flash – zero-shot).....	76
Figure 21: Error-typology Chart (Gemini-2.0-flash - RBP).....	78
Figure 22: Error-typology Chart (DeepSeek-R1 – zero-shot).....	80
Figure 23: Error-typology Chart (DeepSeek-R1 - RBP)	81

Figure 24: Error-typology Chart (Claude-3.5-Sonnet – zero-shot)	83
Figure 25: Error-typology Chart (Claude-3.5-Sonnet - RBP)	85
Figure 26: Heatmap Overview of Quantitative and Qualitative Metric Results of ST-2's TT Outputs.....	86
Figure 27: Error-typology Chart (Google Translate - raw)	88
Figure 28: Error-typology Chart (GPT-4 – zero-shot)	92
Figure 29: Error-typology Chart (GPT-4 – RBP).....	96
Figure 30: Error-typology Chart (Gemini-2.0-flash – zero-shot).....	99
Figure 31: Error-typology Chart (Gemini-2.0-flash – RBP)	101
Figure 32: Error-typology Chart (DeepSeek – R1 – zero-shot)	104
Figure 33: Error-typology Chart (DeepSeek – R1 – RBP).....	106
Figure 34: Error-typology Chart (Claude-3.5-Sonnet – zero-shot)	109
Figure 35: Error-typology Chart (Claude-3.5-Sonnet – RBP).....	111
Figure 36: Heatmap Overview of Quantitative and Qualitative Metric Results of ST-3's TT Outputs.....	112

1. INTRODUCTION

“You’re a translator specializing in...”

Could it have been one of the gaps we have been in pursuit of, this entire time?

Neural Machine Translation (NMT) started to emerge and gain importance around 2013-2014 (cf. Kalchbrenner and Blunsom, 2013; Sutskever et al. 2014; and Cho et al. 2014b, as cited in Bahdanau et al. 2016:1) and improved greatly over years, with a notable shift from conventional Statistical Machine Translation (SMT) (cf. Zens et al., 2002; Koehn et al., 2007, as cited in Lyu et al., 2023:1339) to NMT (cf. Cho et al., 2014; Bahdanau et al., 2015, as cited in Lyu et al., 2023:1339) approaches thanks to the developments in the Deep Neural Networks (DNN) (cf. Lyu et al., 2023:1339). Once sequence-to-sequence models (cf. Sutskever et al., 2014:1) powered by recurrent neural networks (RNN) and attention mechanisms (Bahdanau et al., 2014) were introduced, we wouldn’t be quite wrong in saying a milestone was achieved in a historical sense.

Despite this advance in machine translation (MT) field, yet, there was still a feature which later turned out we lacked when we look through today’s perspective: that is, the available NMT systems of the time (for instance: Google Translate) weren’t conversation-based, nor were they multitasking. They were designed to conduct direct translation tasks of text (and sub-sequences of text) rather than generating “creative” responses. In the following phases of mid-2010’s, the introduction of the “transformer architecture” (Vaswani et al., 2017) further shaped NMT.¹ This architecture offered a foundation for many advancements in Natural Language Processing (NLP) field (Babar, 2024)².

¹The work by Vaswani et al. (2017), titled “*Attention is All You Need*,” introduced the transformer architecture which was a foundational step for today’s use of LLM’s in the field of NLP. Unlike earlier models that used recurrence, the transformer uses a method called self-attention. This allows the model to look at all the words in a sentence at the same time and understand how they relate to each other, which makes it better at understanding context.

²<https://medium.com/@zbabar/the-evolution-of-prompt-engineering-3f67d2607473>

With transformers in place, the NLP field started exploring new possibilities. The introduction of BERT (Bidirectional Encoder Representations from Transformers) by Google (Devlin et al., 2019) and GPT-2 (Generative Pre-trained Transformer) by OpenAI that same year (Radford et al., 2019) shifted the focus from sole machine translation, adding more complex tasks such as question answering, text summarization, and language generation (cf. Sharkey and Treleaven, 2024). These models showed their ability to understand and generate text in a relatively more (even highly more) coherent manner across different tasks (Radford et al. 2019; Devlin et al. 2019). They relied on pre-training³ on large amounts of text data (cf. Radford et al. 2019; Devlin et al. 2019), and this allowed these models to understand language patterns. Thus, they became more functional than conventional NMT systems.

A greater breakthrough happened with GPT-3 in 2020 (Brown et al., 2020). GPT-3 was such an innovative model that it could achieve range of tasks including engaging in question-answering, machine translation, and cloze tasks depending on specific textual needs of the user (i.e.: literary works, informative texts etc.) (cf. Brown et al., 2020:1). The model put forth a rise in its capabilities through its hundreds of billions of parameters, mainly thanks to its ability to generate language and reason contextually (ibid.). It wouldn't be wrong to argue that this offered future potential for human-machine interaction. This paradigm shift was the onset of the method called "prompting" (cf. Babar, 2024).

Prompting is an activity where the user offers an input and demands a large language model (LLM) -with simpler words: an artificial intelligence (AI) model- to provide an output (this could be a textual, visual, even audiovisual output). This process, simply called "prompting", is based on a flow in which the user is essentially building a communicative setting (cf. Jakobson, 1960); positioning themselves as the "sender", the model as the "receiver", preferred chatbot (e.g: OpenAI's GPT-4 model) as the "channel", prompt itself as the "message", and the desired output as the potential "feedback".

³ "Pre-training" refers to the process in which a machine learning model is trained on a large *dataset* of text data before being tailored for a specific task, such as translation, summarization, or question answering.

LLMs began to be tested and constantly improved, thus, became widespread. As they were being constantly improved, they developed the ability to multitask and engage in dynamic, context-aware dialogue with the user. Unlike conventional NMT systems, which simply translated sequences/units between source and target languages, LLMs like GPT-3 could translate, but they could also answer questions and/or respond to prompts, thus, for instance, revise translations when requested. The way machines interacted with human language evolved drastically, upon the introduction of the transformer architecture being followed by development of LLMs like GPT.

With this, when we refer back to the question that I posed at the beginning: “You’re a translator specializing in...” Could it have been one of the gaps we have been in pursuit of, this entire time? This is an outlook of one of today’s driving forces in the field: prompting. We, today, can even prompt an LLM to perform the task, including translation, with key considerations that we communicate to it, that is to say, one can define their audience, purpose, register, style and etc. for the LLM prior to target text (TT) production. (Those we name “considerations” within the frame of translation are “parameters” when it comes to machine learning (ML). Hence be it translation or any other other task, parameters are the variables humans can adjust).

Like many other tasks of LLMs such as question-answering, image generation, etc., translation tasks today can be performed through these specified considerations (target audience, for instance), that are offered to the model via prompt(s). That is to say, one not only can provide the model with an input text and prompt solely to translate, but also can offer the required considerations the model should be aware of -language pair, audience, purpose, context, register and so forth- and translate a text accordingly.

Speaking of considerations, one would naturally question: how does this work with domain-specific practices then, in terms of translation activities? How to satisfy client needs, for example? How to achieve a context-aware TT for the client? Can one successfully integrate the outputs from prompted AI into a translation workflow? Can we oversee quality metrics such as we do for human translation products?

These are essential questions when it comes to and when we look at fields of specialty like medical texts or pharmaceutical texts domains. Medical and pharmaceutical translation domains are type of domains in which the stakes are high. The reason for this is that it's not only about achieving adequate translation outputs, but is also about ensuring that the translated product is organized accurately and consistently, that instructions in such texts are clear, so that patient safety is not compromised by textual or linguistic errors.

The task of translating medical texts requires not only linguistic competence but also profound understanding of the domain itself, just like any other specialized field that also requires adequate knowledge of language for specific purposes (LSP). Möhn and Pelka (1984:26) define LSP as the range of a language that calls for conceptualisation, the notional designation and the recognition of specialised objects and processes pertaining to a particular domain (cf. Möhn and Pelka, 1984:26, as cited in Vlachopoulos, 2017:426). In this frame, the concept of LSP may require us that, one needs to guide a language model to handle the specific requirements of a medical translation project ensuring that the translated product fits to textual and domain conventions, benefitting from well-designed prompting. This can enable us to build a bridge between general language understanding and specialized domain competence, hence, this would make it possible to produce translations that meet the needs of patients, clients and stakeholders in healthcare as well as translation industry.

Therefore, the questions surrounding how to guide translation models in specialized domains like medicine are essential as we move towards more context sensitive translation workflows. To sum up, prompting has had an influence on how human approaches NMT tasks, especially in specialized domains like medical translation. It has enabled us to set the style, tone, purpose and other considerations of translation projects, referring to the gap that conventional NMT systems struggled with. As prompting (both by design and engineering of it) evolves, so may our ability to generate contextually relevant translations.

1.1 Research Questions

This study will try to address the below questions, which aim to reveal both the quantitative and qualitative aspects of machine translation performance. The study seeks to examine not only the effects of different prompting strategies on translation quality but also the broader implications for domain-specific workflows. By doing so, the paper intends to offer a comprehensive understanding of the strengths, limitations and potential integrations of AI-driven translation.

- 1) When compared with zero-shot prompting, do role-based prompts make a meaningful impact on MT outputs? Do they enhance or optimize the quality of the outputs at all?
- 2) What are domain-specific risks, limitations, pros and cons of each *actuator* in such a comparison?
- 3) Can AI be thoroughly relied on or should it be considered to be integrated into human workflows even after prompted results?
- 4) Are long-rooted conventional NMTs still effective when it comes to domain-specific translation tasks?
- 5) To what extent quantitative data, namely metric scores, are determinant when it comes to quality assessment and comparison between LLMs and NMTs? What is the role of a qualitative framework when combined with quantitative tools?

2. AIM OF THE RESEARCH

This research aims to compare the quality of outputs from role-based (persona) prompted pre-trained artificial intelligence models versus their raw outputs (i.e: zero-shot prompted outputs) versus a simply deployed conventional NMT engine: Google Translate, within the domain of medical translation. By doing so, this study has two specific approaches in terms of prompting types: zero-shot (raw) prompting and role-based prompting. LLM models to be deployed will be approached by both simply prompted to translate input texts, then compared by role-based prompting in order to observe how efficient the role-based outputs are. In role-based prompting, LLMs will be offered certain information including: domain, target language, purpose, audience, register, style.

The outputs of this comparative study are expected to put forth the strengths, weaknesses and limitations of the “actuators” in translation life-cycle. The study will primarily try to find an answer to the question of whether incorporating role-based prompts into chatbots have a notable impact compared to zero-shot prompting, when it comes to machine translation activities in medical texts in English⇔Turkish language pair. To offer an overview, it can be said that all the outputs will be divided into three stages: first stage is raw⁴ Google Translate output, second stage is raw LLM outputs, third is role-based prompted LLM outputs.

Using BLEU⁵ (Bilingual Evaluation Understudy) (Papineni et al., 2002) score and COMET⁶ (Crosslingual Optimized Metric for Evaluation of Translation) (Rei et al., 2020) score, the study will assess performance of the said outputs in both quantitative way, and

⁴ Google’s NMT engine have not been specially trained in the scope of this study. There are studies in the literature which aim to oversee specially trained Google Translate engine, however, no training has been applied on the NMT engine in this study. The study only aims to oversee the outputs from the conventional NMT.

⁵ BLEU is an automatic metric for assessing MT quality. The metric compares machine translations to at least one human reference translation. BLEU assesses how closely a candidate translation matches human translations by calculating the precision of n-grams in the candidate translation that also appear in the reference translations.

⁶ COMET is also a metric for assessing MT quality. It builds on language models like XLM-R and is trained with human feedback, so it can better match how people rate translations (Rei et al., 2020, 1–2).

also, will have a qualitative approach by using a customized error typology framework derived from and particularly motivated by MQM (Multidimensional Quality Metrics.).⁷

Current literature is seen to have a limited amount of works in terms of AI outputs of medical/pharmaceutical texts domain and no particular work in English<>Turkish language pair for this persona-prompting scope. The literature also heavily relies on either quantitative metrics alone or error annotation methodology in a qualitative way, yet, this study assumes that a bridge between the two is essential for a solid oversight of quality. According to Castilho et al., NMT engines exhibit promising automated metric results -when compared to other NMT paradigms- however, when human assessment is involved, the results are not as bright (cf. Castilho et al., 2017:117).

⁷ <https://themqm.org>

3. DEFINING THE RESEARCH AREA

This section and its subsections will attempt to elucidate on the core concepts that construct the main research areas of this thesis: prompting LLMs, zero-shot prompting, role-based/persona prompting, medical texts as a domain, and importance of medical texts.

3.1 Prompting LLMs

Prompting refers to the practice of providing inputs called “prompts” to guide a large language model’s outputs, namely responses or behaviour. A prompt can be a range of inputs from simple instructions to questions that intend to elicit answers. When one asks, for instance, “What is the capital of Sweden?” to an LLM, that is a question prompt, and when the answer to come from the LLM as “The capital of Sweden is Stockholm” is considered an output. This is a simple Q&A (question-answer) task in the NLP field. Or, when a user offers an LLM the input: “List 10 biggest pharma companies by revenue” that is an instruction-based prompt by its nature as it commands the model to perform a certain task by instructing it. There are various prompt types in the context of NLP and depending on the task and domain: zero-shot prompting, few shot prompting, chain-of-thought prompting, thread of thought prompting, and so forth (Sahoo et al., 2025).

Besides the types of prompts, in academic works, one can encounter a specialized practice, which is prompt design and engineering -one of the sub-categories of which is types of prompts- that is a practice focusing on utilizing prompts in different efficient ways with the concern of making the most out of the models. Prompt engineering and design is a process of developing pieces of natural language text that act as inputs to LLMs (cf. Desmond and Brachman, 2024:1). Researchers investigate prompt engineering and design in two pillars: on one hand, a set of technical strategies that aim to optimize performance of LLMs, and on the other hand, how users interact with prompts, from a human-machine interaction perspective (cf. *ibid.*).

3.1.1 Zero-shot Prompting

Zero-shot prompting can be considered the core of prompts, meaning the type of prompt which the majority of the ordinary users with no advanced NLP skills who aim to benefit from LLMs resort to. It is the type of prompt which only instructs a model with textual description of the task to be carried out by it (Scius-Bertrand et al., 2024:2) without any preliminary examples provided that the model is to learn from (Vöge et al., 2024:2). In this type of prompting, the model solely relies on its general knowledge that it has been pre-trained with (cf. Sahoo et al., 2025:2)

In the context of machine translation, this applies to practices where a user only provides the source text (ST) to the model and instructs it to produce a TT, without providing any particular reference text, translation memory, style guide, persona prompt, or any of the instruments which a translator resorts to, in the age of translation technologies, thus, the practices that the majority of the ordinary users of AI models will exercise in their daily needs of translation activities, which is under scope of this very particular paper. In this study, this type of prompt will also be referred to as “basic” and “raw” prompt.

3.1.2 Role-based/Persona Prompting

Persona, i.e., role-based prompting, refers to the act of providing input as assigning a large language model a specific role to adopt the role and carry the task (i.e: translation) by adhering to the assigned role specified in the prompt. For instance, one can assign a model a role through such input: “You are a technical writer specialized in the context of instructions manuals. Explain 3 rules of technical writing in instructions manuals” and then expect the model to generate a response accordingly.

Through this method, the output of the task to be carried out is aimed to be enhanced or optimized via such prompting, compared to basic prompting. This is a method also applied in today’s AI-driven MT practices. A user can specify the translator task to an LLM with

desired considerations (even purpose of the text, audience, preferred tone, register, etc.) and desired field of specialty (such as legal, medical, technical etc.).

3.2 Medical Texts as a Domain

Prior to communicating in-depth knowledge pertaining to medical translation, I would first like to invite the readers to have a general overview (so that I can offer a foundation before introduction of further insights) into the domain of medical texts. According to Göktürk, being aware of text types' conventions holds significance in terms of communication to take place (cf. Göktürk, 1994:35). Göktürk adds to his words:

In reality, the concept of text type influences communication in three aspects: first is the indicators to reveal which type of convention the text we are dealing with belongs to. Second, is evoking a certain state of expectation, and as for the third, the shaping of our perception of a text.⁸ (ibid.)

He also argues that comprehensive knowledge on text type conventions is of benefit in terms of identifying the correct translation method to apply on different texts (cf. Göktürk, 1994:36). As for medical domain, according to Göktürk, when, for instance, translating a patient information leaflet of a newly launched pharmaceutical drug, it is essential to focus on that the message of the text is perceived in the easiest way, rather than reflecting the structural features predetermined by conventions (cf. Göktürk, 1994:37).

The medical language to-date, at least in the Western world, often incorporates a hybridity of Greek, Latin (cf. Fischbach, 1986:20) and English (cf. Açıkalın, 1995:26, as cited in Çetintaş Sönmez, 2011:17) terminology. This specialized terminology has a key position in terms of accurate communication in the medical field, since it enables precision and clarity in conveying complex medical concepts. We can also see the emphasis on the importance of absorbing this terminology well, as it serves as a bridge between medical professionals and translators, allowing for effective communication in clinical settings. This

⁸ Translation is my own.

terminological density of Greek and Latin terms require translators to be aware of prefixes and suffixes (cf. Oğuz, 2013:128), alongside eponyms, abbreviations, acronyms, symbols, pronunciation of the terms (cf. Erten, 2016: 228-229, as cited in Günay Köprülü, 2017:17). Lay readers, who are not professionals in the field hence not aware of such terminology, prefer communicating through mundane (daily) counterparts of such terms in their spoken language (cf. Günay Köprülü, 2017:12).

The medical language today exhibits unique linguistic characteristics that set it apart from general language use. One of the most prominent features is its high terminological density, with specialized terms that account for a remarkable portion of these texts. These terms often follow systematic word formation patterns, using Greek and Latin morphemes in a predictable manner, which helps with term recognition and understanding across different languages -as long as formed properly- (cf. Džuganová, 2013:56).

Having aforementioned general language use, it would be worth drawing attention towards the duality of “Language for Specific Purposes (LSP)” (cf. Dan 2001:149 as cited in Çetintaş Sönmez, E. 2011:13) versus “Language for General Purposes (LGP)” (cf. Picht 1987: 149 as cited in *ibid.*). When we both look at the literature and practice today, medical language is positioned on a more LSP scale, however, parallel texts for communicating the content of those texts to non-professional audiences are then required to be produced.

Then, there emerges the need for use of plain language depending on the communicative setting in which the communication (verbal or written) is taking place. “Plain language” refers to the use of language that is clear, straightforward, and easy for the general public to understand. It avoids jargon and complex terminology, making medical (that is the technical language here) information accessible to patients and their environment (e.g: families). Patient leaflets, in particular, need to be prepared taking into account that these texts need to respond to every potential question in mind of its audience, and should constitute a style aligning with this (cf. Yağbasan and Kurtbaş, 2006:101). To balance these two poles of a communicative setting is a sensitive consideration to keenly follow.

3.3 Importance of Medical Translation

Medical translation, as a domain, has long constituted a great significance for humanity, as it directly relates to services pertaining to healthcare and pharmaceutical industry. Medical translation activities, thus, interact with practices that have direct impact on human health. With the advancement of translation technologies in today's industry, more and more texts within the field have been subject to translation activities.

The practice of medical translation, however, has not emerged recently. From a historical perspective, medical translation may be one of the oldest domains alongside religious translation, since suffering of body and soul have been always the primary focus of us humans (cf. Berghammer, 2006:40). Medical translation activity traces back to earlier ages than modern era, across a vast range of spaces: when one looks first at pre-literary history, Ancient Mesopotamia, a geography to have hosted oldest forms of clay tablets embodying cuneiform writings, is known to have undergone medical translation activities (cf. Montalt et al., 2018:27).

Between 6th and 7th centuries AD, some of the Hippocratic and Galenic texts were translated into Latin in Italy, despite the downgrade of Greek knowledge and medical practice in Europe after the fall of Western Roman Empire (cf. Tracey, 2018:1300). Another remarkable movement of medical translation was the one in the early Middle Ages, when Arabic physicians translated ancient Greek medical texts into Syriac and Arabic (ibid.). Lastly, translations that took place during the Baghdad movement between the 9th and 10th centuries were stored in Moorish Spain which led to further translations into Latin and Castilian in a later period (cf. Tracey, 2018:1301).

Medical translation, therefore, has been fundamentally inherent to the translation systems shaping history ever since pre-historian ages. The (post)modern era, on the other hand, has brought significant advancements in medical translation –just as it has in most other domains– reflecting further technological and cultural changes. In today's world, the demand for timely medical translations has been increasing. Today, translation technologies and instruments, ranging from machine translation and localization softwares to AI

prompting, have contributed to the transformation in the way medical texts are rendered across languages. These tools are assumed to enhance efficiency and accessibility, allowing healthcare professionals, stakeholders and patients to communicate vital information accurately.

Moreover, the status quo of medical translation now has a more hybrid approach, combining human expertise with automated processes. While technology (by technology, I refer to a dual, hand-in-hand reception of MT and AI co-existence as two common and rapidly advancing instruments) provides the industry and its actors with speed, human translators are essentially needed for ensuring (con)text-awareness, domain sensitivity and contextual accuracy, especially in complex medical settings. From translation perspective, this hybrid setting has been important in terms of highlighting the complexities of medical terminology, clinical trials, medical research, and even technical products such as patient leaflets for use of pharmaceutical drugs. As we look at the challenges of today's healthcare, the development of medical translation shows us both the technological progress achieved and also its crucial role in safety of patient health and providing people with fair access to medical information around the world.

4. LITERATURE REVIEW

4.1 Machine Translation (MT) Optimization via Prompting

With the advancements particularly taking place in NLP and particularly in MT field in the last few years, MT optimization via prompt design has gained interest rapidly.

In the literature, several comprehensive studies have been carried out to investigate if and to what extent prompt design and engineering are effective on the quality of the MT outputs of LLMs.

Masaru Yamada (2023) contributes to the field by investigating how translation quality can be impacted when one integrates translation purpose and target audience specifications into prompts. Based on Yamada's study it can be inferred that conventional MT approaches overlook these pre-production considerations, that are well-established in ISO 17100 (cf. Yamada, 2023:195).

By designing prompts that define translation parameters, such as purpose of the translated product alongside audience, Yamada discovers that LLMs like ChatGPT can alter the quality of the MT outputs by industry standards (cf. Yamada, 2023:203). It can also be understood that prompt-driven translation, when combined with suitable prompts, can enhance quality, particularly in marketing and idiomatic expressions, where conventional MT systems often struggle (cf. Yamada, 2023:200).

Yamada benefits from Open AI's word embedding API in order to calculate cosine similarity to evaluate translation outputs. Using OpenAI's text-embedding-ada-002 model, he calculates the cosine similarity between source and target texts to capture their semantic closeness. While this method provides a measure of how closely a translation aligns with the source text, one can acknowledge its limitations: translation quality is not solely determined by lexical or structural similarity. Therefore, he supplements cosine similarity calculations with qualitative human evaluation.

Similar to Yamada's work, Gao et al. (2023) explore how prompt engineering can improve ChatGPT's MT performance across different prompts and big conventional NMT

systems like Google Translate and DeepL. Their study is an assertion of that simple, naive prompts such as basic “*Translate this sentence*” instructions do not fully unleash ChatGPT’s capabilities. In order to address this issue, the study proposes structured translation prompts that explicitly define the translation task, provide domain-specific information, or incorporate linguistic features such as part-of-speech (POS) tags.

In their study, Gao et al., evaluate ChatGPT’s performance. They benefit from standard metrics like BLEU, ChrF++ (Character n-gram F-score) and TER (Translation Edit Rate) scores. The results are compared to those of conventional commercial MT tools. Their findings suggest that ChatGPT outperforms zero-shot prompted translations when structured prompts are applied. Their study reasserts that when prompts, that define translation task clearly and incorporate domain-specific information, are well-designed, the quality of MT using ChatGPT can be remarkably improved.

Lastly, another study conducted on different prompting strategies, that is worth mentioning, is by Zhang et al. (2023). Their study draws on GLM-130B model focusing on MT optimization via different prompting strategies. They rely on different prompt templates and demonstrations (i.e: few-shot examples). In the study, how such factors influence translation performance is analyzed, using the COMET metric. The research notes the importance of structured prompt templates alongside carefully selecting demonstration examples. Gao et al. found that concise prompts, that define source and target languages explicitly, outperform more elaborate prompts.

4.2 Persona Prompting in Large Language Models (LLMs)

Persona prompting, also known as role-based prompting, is an emerging field that is subject to future research that can unfold the capabilities of this prompting approach. To date, it is a field in which relatively limited work has been conducted. Especially for machine translation, there is a notable gap in the literature when it comes to persona prompting’s effect on translation quality and performance. Few studies focus on various tasks to be exercised through role-based prompting, however, there is a great absence of investigation on how

quality outputs this type of prompting can bring forth. More broadly, persona prompting is a process where AI models are provided with explicit instructions to assume specific roles or personas when generating outputs. By this approach, giving clear guidance to the model on the role it should assume, it is presumed to become easier for the model to produce outputs tailored to particular contexts, styles, or areas of expertise.

Recent research by Zheng et al. (2024) investigates the effects of persona prompting on large language models' performance in general. They compile a list of 162 roles consisting of 6 types of interpersonal relationships and 8 domains of expertise (2024:15126). Their work reveals that incorporating specific personas into system prompts does not necessarily increase model accuracy on various tasks. However, further analyses by the authors demonstrate that factors such as the gender, type, and domain of the persona could impact prediction accuracy.

The authors also note that selecting the best persona for each question could remarkably improve performance, however, automatically identifying the most effective persona remains challenging and often lead to results comparable to random selection (cf. Zheng et al., 2024:15126). The results emphasize the variability and unpredictability in the impact of individual personas because in some models the persona prompting strategy can do worse than random selection (cf. Zheng et al., 2024:15134). Thus, the authors add to this unpredictability by noting that it is currently uncertain whether role-playing would affect LLMs' capability to perform general language tasks (ibid.).

4.3 Generative AI and Machine Translation in Medical Domain

As for many other domains, other areas of specialty, generative AI's influence on medical translation activities as well have gained much attention among researchers upon the latest advancements in LLMs and NMT. Some researches focus on LLM's effectiveness as being the primary actuator in MT workflows, while some studies (Genovese et al., 2024:11) suggest that AI-integrated workflows are more of supplementary practices when human translators are not accessible. Nonetheless, the rapid advancements achieved in the NLP field do not

remain unseen. Chatbots, for instance, can now receive user-uploaded documents as source texts, perform voice-recognition and produce outputs accordingly, and function in a versatility of tasks. As for AI models released prior to these advancements, however, the scope of some researches had its limits.

Canbaz (2024), in her work, focuses on ChatGPT's GPT-3 model's performance and position in the pre-, during and post-translation activities in medical texts domain and questions how essential an involvement by a human translator is. Canbaz notes that the model lacking the feature to receive document uploads (2024:53) by users leads to the need to manually extract the ST input and paste into the chatbot. This had to be done via pre-editing the input due to data processing restrictions by the LLM-provider (cf. Canbaz, 2024:147). Canbaz resorts to BLEU metric in her study, yet, she notes that since BLEU metric is a limited tool in terms of contextuality, linguistics and comprehensibility, it is not fully reliable to be able to determine whether the outputs are adequate or acceptable translations (cf. Canbaz, 2024:145).

Her work suggests that, in informative texts such as medical texts, the TT being not textually fluent yet preserving the information conveyed can be sufficient for the target audience, however, for lay readers, pre-editing and post-editing practices will enhance textual fluency and clarity (cf. Canbaz: 2024:144). Her research overall argues that, in medical domain, GPT-3 is a model that is inadequate in terms of translation quality and performance and requires human involvement due to its poor intertextual cohesion, occasionally disobeying the prompts, and altering the meaning in sentence-based translations (cf. Canbaz, 2024:148-149).

Another work in the literature that I would like to mention with respect to conceptualization of LLMs as translator chatbots is Alimen's study conducted on the chatbot GPT 3.5's influence on translation practices and translator training (Alimen, 2023:1532), when translation briefs are incorporated into the prompts (cf. Nord, 1997:30, as cited in Alimen, 2023:1532). In her experimental research, Alimen evaluates the quality of the translated product, the research material of which is a patient information brochure extracted from a hospital website, based on the error-typology categories: accuracy, fluency and

terminology, that are included in MQM (cf. Alimen, 2023:1537). Her study asserts that the need for clear and correct prompting is essential to achieve quality outputs, as the language model unquestioningly accepts the prompts offered, and when the model is misguided, the outputs it produce can conflict with the purpose of the translation commission (Alimen, 2023:1544). She also concludes by motivating future research by students of translation discipline to further investigate what kind of prompts are effective when it comes to operate ChatGPT in a more functional fashion (cf. Alimen, 2023:1545).

In the literature, there are studies that focus on domain-specific fine-tuned language models in medical domain for NMT purposes. Han et al. (2024:1) carried out a research where they investigate MT practices on clinical texts using deep learning methods such as transformers-based structures. They aimed to conduct -same as in this research- both a quantitative and a qualitative framework (2024:7) to evaluate translations. Their work first used standard automated metrics to assess the performance of three fine-tuned multilingual models: Marian, NNLB and WMT21fb, across the sub-tasks of clinical case reports, clinical terms and ontological concepts (2024:1).

On one hand, they employed metrics such as BLEU, METEOR, ROUGE, COMET (2024:2) to perform quantitative evaluation and on the other hand, they deployed an expert-based post-editing quality evaluation metric, -named “HOPE” by Gladkoff and Han⁹- for which they recruited human evaluators (who have backgrounds as professional translators, linguists and biomedical researchers) (cf. Han et al., 2024:7). This qualitative evaluation metric has predefined error types and penalty points according to the severity of the errors, same as this paper introduces. In their work the findings revealed that a small-sized pre-trained language model, Marian, outperformed two extra LLMs, NNLB and WMT21fb, in terms of clinical domain fine-tuning (2024:1).

As for takeaways of the study, it can be inferred that model size does not necessarily guarantee higher quality outputs. Furthermore, despite automated metrics such as BLEU,

⁹ Gladkoff S, Han L. HOPE: a task-oriented and human-centric evaluation framework using professional post-editing towards more effective MT evaluation. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association (2022). p. 13–21.

COMET, etc. being useful, they do not necessarily reflect translation quality. Hence, same as Canbaz (2024), Han et al. advocate human-in-the-loop translation workflows since human involvement in terms of quality evaluation becomes essential (cf. Han et al., 2024:11-12). This current study supports the proposition of Han et al., in the sense that automated metrics alone cannot have a final say on the quality of MT outputs, thus, a qualitative approach becomes a need for validation purposes.



5. METHODOLOGY

The methodology for this experimental study will draw on two pillars: pre-production and production of translated text. In the first pillar: the three source texts will be presented, relevant prompt designs and other preliminary activities will be elucidated. Initially, the first source text will be presented, followed by the reference text and prompts pertaining to it. Subsequently, the same cycle of steps for the second and third source texts will take place. In addition, the tools deployed, namely, the actuators to perform translation act will be subject to discussion. In the second pillar, both quantitative and qualitative results will be elaborated on, with detailed commentaries following them.

This approach combines two ways of looking at translation quality: using numbers to measure performance and adding a deeper analysis to see if numbers alone tell the whole story. While metrics like BLEU score give precise and measurable results, they don't always necessarily reveal how well the translation fits its purpose or works in the target language conventions. Adding a qualitative framework, such as an error-typology framework, helps us to check if the translation keeps the intended meaning and works well for the new audience. By using both methods together, we can get a clearer and more complete picture of the translation's strengths and weaknesses (cf. Munday, 2016:294). This way, we balance accuracy with text's function and impact.

5.1 First Pillar: Pre-production Process of Translated Product

This section, with its subsections, is intended to guide the reader through one of the very essential stages of a translation life-cycle: pre-production process. Drawing on and motivated by the international standard ISO 17100:2015, this pillar is designed to define the preliminary activities I carried out prior to the production process of the TT outputs that are to be examined within the scope of this paper.

Pre-production process is an essential stage, in the sense that it is where the TSP's (translation service provider), LSP's (language service providers), translation professionals

define the structure of the overall production process. They don't only review client needs and expectations to be met, but they also assess source text and define tools to be in place for the candidate translation(s) along with their efficiency. Prior to production process, according to ISO 17100:2015 (ISO, 2015:7), TSP's should implement practices that should conduct the necessary project feasibility and handle the procedures prior to entering into an agreement with a client, such as preparing quotations or determining whether the necessary tools for a delivery of efficient translation product are available.

"Feasibility" is a vital concept because in any business operation, it is a preliminary process that helps service providers and receivers determine whether their mutual expectations will be met with efficiency. This applies to translation activities as well, from source text selection to whole project setup.

As, within the scope of the study, we don't have a client that requires us to deliver a certain product, we will have a general oversight on the source texts and target texts, with the prompts and specifications determined in the prompts. I will, however, elaborate on the basic feasibility procedures I carried out alongside various preliminary actions I took.

5.1.1 ST-1

Before delving into the production process, it's important to have an overview of the source text. This subsection will comment on the rationale of ST selection. Below is the first ST that is subject to this experimental research.

Table 1: ST-1

Randomised phase 3 study vs. dacarbazine (CA209066)
--

The safety and efficacy of nivolumab 3 mg/kg for the treatment of advanced (unresectable or metastatic) melanoma were evaluated in a phase 3, randomised, double-blind study (CA209066). The study included adult patients (18 years or older) with confirmed, treatment-naïve, Stage III or IV BRAF wild-type melanoma and an Eastern Cooperative Oncology Group (ECOG) performance-status score of 0 or 1. Patients

Table 1 - continue

with active autoimmune disease, ocular melanoma, or active brain or leptomeningeal metastases were excluded from the study.

This is an extract from SmPC (summary of product characteristics) document of an oncological pharmaceutical drug named “Opdivo”¹⁰, used in treatment of various types of cancer. Its active substance is called “nivolumab”. The drug is manufactured by the pharmaceutical company Bristol Myers Squibb.¹¹ The extract mentions the active substance being evaluated in a previous clinical trial and patient population of the participant patients.

The text is a technical and informative text by function. As the intended audience is healthcare professionals, it embodies LSP usage, with coinage terms from Greek and Latin languages of medicine. Its register is formal. According to Article 4.6.3.1. of ISO 17100:2015, TSP’s should conduct a source language content analysis during the course of their pre-production activities.

There are a few reasons for selecting this extract. In documents such as SmPC, their locally adapted versions, namely target texts, can have relative differences compared to the original text. For instance, there can be omitted or added bullets, sentences, even paragraphs. Therefore, this hinders a good reference text selection as well. As there is BLEU metric to be used within the quantitative frame of this study, a reference text with more direct translation was needed. Unlike some paragraphs, this extract was translated without any additions or omissions in the Turkish SmPC document. Thus, this was a solid reason for selecting the extract.

Another reason lies in the nature of the text. It is filled with numeric details, parentheses, and hybrid terms of pharmaceutical and clinical research terminology, which can be challenging for LLMs. LLMs often tend to favor one domain over the other, leading to potential inconsistencies. Additionally, LLMs can produce critical errors even with widely

¹⁰ <https://www.opdivo.com>

¹¹ <https://www.bms.com>

recognized clinical trial terminology. For instance, the term “randomized” may be incorrectly rendered as “rastgele” in Turkish, rather than the common-ground term “randomize”, as will be further elaborated during the post-production analysis. Thus, this extract presented a demanding test case, enabling me to examine how prompt optimization within the study could minimize such errors.

One final reason worth mentioning is the intended audience. I selected two source texts for this study, one is designed for general audience and the other is for healthcare professionals. This approach allows the audience of this paper to gain insights into the varying risks, limitations, and challenges associated with both target groups. This extract in question, as it is from a SmPC document, is designed to address to healthcare professionals.

Source text selection, therefore, is as important as prompt design because it has a potential impact on the performance and reliability of LLMs. A carefully chosen source text not only ensures a more accurate evaluation of translation quality but also enables us to understand how textual, contextual and linguistic challenges influence machine-generated outputs.

5.1.1.1 Reference Text-1

Below is a reference text from the SmPC document’s Turkish translated text. As it is an official translation of the ST, also published by the relevant regulatory authority in Turkey, we will take this as reference when performing translation quality evaluation (TQE) for the TT outputs. When both BLEU metrics are calculated and error-typology is performed, TT outputs will be evaluated on how “close” they are to this reference text. The degree of similarity between the TT outputs and this reference text will help determine the overall translation quality and highlight specific areas where improvements are necessary.

Table 2: Reference Text for ST-1

Dakarbazin karşısında randomize faz 3 çalışma (CA209066)
--

Table 2 - continue

Nivolumab 3 mg/kg'ın ilerlemiş (rezeke edilemeyen veya metastatik) melanom tedavisi için güvenliliği ve etkililiği bir faz 3, randomize, çift kör çalışmada (CA209066) değerlendirilmiştir. Çalışmaya doğrudan, tedavi uygulanmamış, Evre III veya IV BRAF vahşi tip melanomu olan ve ECOG performans durumu 0 veya 1 olan yetişkin hastalar (18 yaş ve üzeri) dahil edilmiştir. Aktif otoimmün hastalığı, oküler melanomu veya aktif beyin ya da leptomeningeal metastazları bulunan hastalar çalışmaya alınmamıştır.

5.1.1.2 Zero-shot (Raw) Prompt

Below is the first prompt to provide as input to each LLM for the ST-1:

Table 3: Zero-shot Prompt Provided to LLMs for Translation of ST-1

Translate this text from English to Turkish.

The rationale of choosing such prompt is to evaluate what kind of an output LLMs will generate against more structured and specified prompts. Therefore, I decided that such raw prompt was needed for my methodology.

5.1.1.3 Role-based Prompt

Below is the details of the prompt in which I prompted each LLM to simulate as an oncology translator:

Table 4: Role-based Prompt Provided to LLMs for Translation of ST-1

You are a medical translator specialized in oncology field. Translate the below-given [source text] adhering to the following considerations:

Purpose: Translation of an extract from a Summary of Product Characteristics document about an oncological pharmaceutical drug.

Table 4 - continue

Audience: Healthcare professionals
Style: Technical
Register: Formal
Style guide: No specific style guide to be provided however please conform to up-to-date clinical trial terminology in Turkish, preserving accuracy and consistency.

In the context of this study, by role-based prompting, I assigned the model a defined role, such as that of a medical translator, aiming for generation of more contextually accurate and terminologically consistent translations. In translation activities, this approach helps mitigate common errors in specialized domains, such as medical and pharmaceutical texts, by providing the model with a clear framework and expectations. For example, by instructing the model to “...Translate the below-given [source text] adhering to the following considerations:...” it was aimed to potentially lead to improved accuracy aligning with domain-specific conventions.

Following were defined in alignment with that of ST: purpose, audience, style and register. A style guide is not in place since this is not within the scope of this study. Our main focus is performance of LLMs upon prompting, thus, it was stated to each LLM under this prompt to cope with up-to-date terminology of clinical trials.

5.1.2 ST-2

Below is the second source text that is within the scope of this experimental study:

Table 5: ST-2

Breast-feeding
Do not breast-feed while you are taking this medicine. It is not known whether the active substance in Venclyxto passes into breast milk.

Table 5 - continue

Fertility

Based on findings in animals, Venclyxto may cause male infertility (low or no sperm count). This may affect your ability to father a child. Ask your doctor for advice on sperm storage before starting treatment with Venclyxto.

Driving and using machines

You may feel tired or dizzy after taking Venclyxto, which may affect your ability to drive or use tools or machines.

This is an extract from the PIL (patient information leaflet) document of an oncological pharmaceutical drug, named “Venclyxto”¹², manufactured by the pharmaceutical company AbbVie¹³. Just like ST-1, the text is a technical and informative text by function. The intended audience for this, however, is general audience. Nevertheless, it embodies a hybrid of LSP and LGP. Its register is formal.

I selected this text mainly because I wanted to put different types of texts, audiences of which are different from one another under the scope of this study. The former one addresses to more of technical experts and this one is intended for more laymen.

5.1.2.1 Reference Text-2

Below is the reference target text against the ST-2.

¹² <https://en.wikipedia.org/wiki/Venetoclax>

¹³ <https://www.abbvie.com/>

Table 6: Pre-edited Reference Text for ST-2

Emzirme

Bu ilacı kullanırken emzirmeyiniz. VENCLYXTO’nun etkin maddesinin anne sütüne geçip geçmediği bilinmemektedir.

Fertilite (üreme yeteneği)

Hayvan çalışmaları bulgularına göre, VENCLYXTO erkek infertilitesine (sperm sayısında düşme ya da sperm yokluğu) neden olabilir. Bu durum baba olma yetinizi etkileyebilir. VENCLYXTO tedavisine başlamadan önce sperm depolama hakkında doktorunuza danışınız.

Araç ve makine kullanımı

VENCLYXTO kullandıktan sonra, araç ve makine kullanımı kabiliyetinizi etkileyebilecek yorgunluk veya baş dönmesi hissedebilirsiniz.

5.1.2.2 Zero-shot (Raw) Prompt

Just as in the zero-shot prompt for ST-1 translation, this zero-shot prompt is designed the same way so that we can observe raw outputs before the role-based prompts. Below is the zero-shot prompt:

Table 7: Zero-shot Prompt Provided to LLMs for Translation of ST-2

Translate this text from English to Turkish.

5.1.2.3 Role-based Prompt

Below is the role-based prompt for the ST-2 translation:

Table 8: Role-based Prompt Provided to LLMs for Translation of ST-2

You are a medical translator specialized in oncology field. Translate the below-given [source text] adhering to the following considerations:

Purpose: Translation of an extract from a Patient Information Leaflet document about an oncological pharmaceutical drug.

Audience: General audience

Style: Technical

Register: Formal

Style guide: No specific style guide to be provided however please conform to medical/pharmaceutical terminology in Turkish, preserving accuracy and consistency.

The difference from the previous role-based prompt for ST-1 is that, it is defined in the purpose that this is an extract from a PIL document rather than a SmPC document. Plus, the audience is this time general audience instead of healthcare professionals, as PIL documents address to general public unlike SmPC documents. Translations of PIL documents require a more functional approach that focuses on clarity, simplicity, and reader comprehension while preserving accuracy and technical style used. One of the key goals is not just to convey medical information but to ensure that non-professional audience can easily understand the information provided.

Every translation aims to fulfill a specific communicative purpose for its target audience. By analyzing the characteristics, needs, and expectations of the intended audience, translators can produce a text that effectively meets those requirements. (cf. Nord, 2006:133)

5.1.3. ST-3

Below is the third and last source text within the scope of this study.

Table 9: ST-3

STEP 2

Place the following items on a clean, flat surface:

- 1 pre-filled pen
- 1 alcohol pad (not included in the carton)
- 1 cotton ball or gauze pad (not included in the carton)
- special disposal container (not included in the carton)

Wash and dry your hands.

STEP 3

Choose from these 3 areas to inject:

- front of left thigh
- front of right thigh
- your belly (abdomen) at least 5 cm from your belly button (navel)

Before the injection, wipe where you will inject in a circular motion with an alcohol pad.

- **Do not** touch or blow on the injection site after it is cleaned. Allow the skin to dry before injecting
- **Do not** inject through clothes
- **Do not** inject into skin that is sore, bruised, red, hard, scarred, or has stretch marks
- **Do not** inject into areas affected by psoriasis

5.1.3.1 Reference Text-3

Table 10: Reference Text for ST-3

ADIM 2

Aşağıdaki malzemeleri temiz ve düz bir yüzey üzerine yerleştiriniz:

- 1 adet kullanıma hazır kalem
- 1 adet alkollü ped (kutuya dahil değildir)
- 1 adet pamuk yumağı veya gazlı bez (kutuya dahil değildir)
- Özel atık kabı (kutuya dahil değildir)

Ellerinizi yıkayıp kurulayınız.

ADIM 3

Enjeksiyon için şu 3 bölgeden birini seçiniz:

- sol uyluğun önü
- sağ uyluğun önü
- göbek deliğinin en az 5 cm uzağında olmak üzere göbek bölgesi (karın)

Enjeksiyondan önce, enjeksiyon yapılacak yeri bir alkollü ped ile dairesel hareketlerle siliniz.

- Temizledikten sonra enjeksiyon bölgesine **dokunmayınız** ya da **üfleme** yapmayınız.

Enjeksiyondan önce derinizin kurummasını bekleyiniz.

- Kıyafetlerin üstünden enjeksiyon **yapmayınız**.
- Ağrıyan, morarmış, kızarmış, sertleşmiş, yaralı veya çatlak izleri olan deri üzerine enjeksiyon **yapmayınız**.
- Sedef hastalığından etkilenmiş bölgelere enjeksiyon **yapmayınız**.

5.1.3.2 Zero-shot (Raw) Prompt

Table 11: Zero-shot Prompt Provided to LLMs for Translation of ST-3

Translate this text from English to Turkish.

5.1.3.3 Role-based Prompt

Table 12: Role-based Prompt Provided to LLMs for Translation of ST-3

You are a medical translator specialized in dermatology field. Translate the below-given [source text] adhering to the following considerations:

Purpose: Translation of an extract from a Patient Information Leaflet document about a dermatologic pharmaceutical product.

Audience: General audience

Style: Technical

Register: Formal

Style guide: No specific style guide to be provided however please conform to medical/pharmaceutical terminology in Turkish, preserving accuracy and consistency.

5.1.4 Actuators Deployed

ISO 17100:2015 also suggests delegation of roles in a project team prior to translation process, basically within pre-production process. As this paper doesn't work with humans, the actuators of the experiment will be one NMT engine and 4 widespread LLM's, including the following:

- Google Translate
- GPT-4 by OpenAI
- Gemini-2.0-flash by Google
- R1 by DeepSeek
- 3.5-Sonnet by Claude

5.1.5 Qualitative Framework: TQE (Translation Quality Evaluation) Penalty Score Tool

It is one of the emerging focus areas of translators in their profession to deal with human quality evaluation on MT, which has an association with post-editing tasks (Öner and Öner Bulut, 2021:101). This helps in facilitating both capturing different errors (hence having a strong oversight of MT-driven processes) in a MT-produced text and reducing the effort spent when exercising post-editing activities (cf. Öner and Öner Bulut, 2021:101) so that the MT outputs can be altered more effectively. As mentioned at an earlier stage, I have created a customized TQE tool to have a comprehensive overview of quality of MT outputs.

The primary rationale of this is that quantitative metrics alone are not sufficient to have a say on final quality of an MT output since they heavily rely on unit matches between reference and candidate texts. Metric evaluation is an automated process and for translation activities, a solid overview of quality should be in place as well. That is to say, numbers can have a say to some extent, yet, especially in translation of texts like medical texts, critical thinking becomes a must rather than solely relying on numbers that are derived from computational formulas.

Therefore, to adopt critical thinking over the MT outputs, I have decided that there is a need for contribution by an error-typology tool. Motivated by MQM, I created my own error-typology spreadsheet alongside a legend, scoring matrix and instructions on how to use the TQE penalty tool. In the following sections, the tool and its said components will be presented and elaborated on.

5.1.5.1 TQE Spreadsheet

Below is the template chart, namely the TQE spreadsheet, which I use to enter the count of error instances.

A	B	C	D	E	F	G
Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use				0	0
Terminology	Incorrect term use				0	0
Accuracy	Mistranslation				0	0
Accuracy	Overtranslation				0	0
Accuracy	Undertranslation				0	0
Accuracy	Addition				0	0
Accuracy	Omission				0	0
Accuracy	Untranslated				0	0
Linguistic Conver	Grammar				0	0
Linguistic Conver	Spelling				0	0
Syntax	Syntactical errors				0	0
Style & Clarity	Unnatural Phrasing				0	0
Style & Clarity	Poor Readability				0	0
Custom	Layout, character & other				0	0
		Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
		0	0	0	0	0

TQE Spreadsheet
Legend
Scoring Matrix
Instructions
+

Figure 1: Screenshot of TQE Spreadsheet Template

How this tool works is a two-step process. In the first place, prior to entering any data into this spreadsheet, I conduct a manual error analysis on any MT output. I identify all the errors and note down the error type and severity (minor, major or critical), and the count of instances, meaning how many times a specific error occurred in the same output. Having completed my manual analysis, I then go to this tool ready to enter data.

Prior to explaining how I enter data, I feel the need to focus on what this tool consists of. This spreadsheet consists of 7 main columns (Columns A to G). Column A is the main category of the quality issue, namely, the error category. Column B is the sub-category, meaning the subtype of the quality issue. Columns C-D-E are the severity levels of the error: minor, major or critical. For instance, if a mistranslation issue has been identified with major severity, and if there are 2 mistranslation errors in the candidate text, I enter “2” in the following field: *D4* (because Column D is for major and Row 4 is for mistranslation). Basically I enter the count of an error instance in under the correct severity column.

I constructed this Excel sheet with formulas: minor errors are assigned 1 point, major errors are assigned 3 points, and critical errors are assigned 5 points. When all the errors

identified are entered on the spreadsheet, it automatically calculates the sum of all the errors identified and calculates both each error's penalty score and the final total penalty score, to which I refer as "TQE penalty score".

Let us go over a TQE analysis scenario: imagine we have a TT output that has the following:

- Undertranslation error → 2x instances → minor severity
- Omission error → 1x instance → major severity
- Poor readability error → 1x instance → critical severity

When I enter the numbers of instances, the following result appears:

	A	B	C	D	E	F	G
1	Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
2	Terminology	Inconsistent use				0	0
3	Terminology	Incorrect term use				0	0
4	Accuracy	Mistranslation				0	0
5	Accuracy	Overtranslation				0	0
6	Accuracy	Undertranslation	2			2	2
7	Accuracy	Addition				0	0
8	Accuracy	Omission		1		1	3
9	Accuracy	Untranslated				0	0
10	Linguistic Conver	Grammar				0	0
11	Linguistic Conver	Spelling				0	0
12	Syntax	Syntactical errors				0	0
13	Style & Clarity	Unnatural Phrasing				0	0
14	Style & Clarity	Poor Readability			1	1	5
15	Custom	Layout, character & other				0	0
16							
17			Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
18			2	1	1	4	10

Figure 2: TQE Penalty Result of an Example-based Scenario

As there are formulas I embedded in the Excel sheet, when one enters the count of instances, Excel automatically calculates total errors in each severity, and also the final penalty score based on the severity points assigned. The total penalty score of this imaginary output is 10. It is worth mentioning that as this is an error-typology tool, the higher the total point is, the worse it is as a high TQE penalty point is a negative instance.

5.1.5.2 Error-typology Legend

I created a legend where each quality issue and error type are defined and so that when conducting a translation quality analysis, I can refer to the legend to determine the correct type of error instance. Following is a screenshot of the TQE spreadsheet's legend:

Quality Issues and Detailed Error Types Legend:				
1. Terminology				
• Inconsistent use: Use of various terms for the same concept to be signified in cases where consistency should be maintained.				
• Incorrect term use: Use of a term that wouldn't be an appropriate fit for a specific domain concept. A conceptual mismatch.				
2. Accuracy				
• Mistranslation: Error arising when the TT content misrepresents/falsifies the source content or any piece of information in the source content				
• Overtranslation: An inappropriately more specific translation of source content than intended.				
• Undertranslation: An inappropriately less specific translation of source content than intended.				
• Addition: An error in the case of a piece of information or a part of speech that is not included in the source content yet present in TT.				
• Omission: An error in the case of a piece of information or a part of speech that is included in the source content yet missing in TT.				
• Untranslated: The type of error in which a sequence or a single word is in appropriately left in the source language, rather than translated.				
3. Linguistic Conventions				
• Grammar: An error where there is a violation of grammatical rules in the target language.				
• Spelling: When there is a misspelling issue in TT output.				
4. Syntax				
• Syntactical Errors: Error occurring when there is syntax-level intricacies.				
5. Style & Clarity				
• Unnatural Phrasing: Use of phrases or sentences that are unnatural in the target language.				
• Poor Readability: Text structure or vocabulary that makes reading and understanding difficult.				
6. Custom				
• Layout, character & other: Errors that are related to layout, character or other types of errors not adhering to TT conventions				
►	TQE Spreadsheet	Legend	Scoring Matrix	Instructions +

Figure 3: Screenshot of TQE Spreadsheet Legend

5.1.5.3 Scoring Matrix

Another tab in the Excel tool that can be referred to is the scoring matrix. It explains error severity levels and the rationale of the severity levels. Following is a screenshot of the scoring matrix:

Scoring Matrix:									
Severity Levels:									
Critical (5 points): Errors that could lead to serious misunderstanding, misuse of the product (medication), or health risks on user's end.									
Major (3 points): Errors that complicate the meaning or cause significant ambiguity or confusion.									
Minor (1 point): Errors that are noticeable but do not significantly hinder understanding.									
IMPORTANT: Threshold for acceptable translation is 16 points.									
16> is considered low-quality translation.									
(Rationale of this threshold is ---> 1 critical, 2 major, 5 minor error instances MAXIMUM are tolerated.)									
◀ ▶ TQE Spreadsheet Legend Scoring Matrix Instructions +									

Figure 4: Screenshot of Scoring Matrix

In the matrix, it is also noted that the threshold for an “acceptable” translation is maximum 16 TQE points. The rationale behind this is 1 critical, 2 major and 5 minor error instance maximum are tolerated, which equal to 16 points (TQE score that is 16> is considered low-quality translation).

5.1.5.4 Instructions for Use for TQE Tool

For anyone that is new to this tool, there are instructions available on how to use the tool. Following is a screenshot of the instructions tab:

◀ ▶ TQE Spreadsheet Legend Scoring Matrix Instructions									
Instructions for Use: • Identify and record instances: Count each instance where an error occurs. • Assign penalty: Use the provided definitions to determine the penalty (severity). • Calculate impact: Multiply the number of instances by the points assigned to each severity level to calculate the total impact score for each error type. • Sum Total Points: Add up the total points for each category to determine the overall translation quality.									

Figure 5: Screenshot of Instructions

5.2 Second Pillar: Production Process of Translated Product

As there is no project team in this experiment, this pillar will not focus on Article 5.2. of ISO 17100:2015 -which is “Translation project service management” during production process- although, it will focus directly on Article 5.3.1 which is “Translation” under “Translation process” which defines the aspects to conform to during the translation process. Commentary section will be more parallel to the international standard’s relevant section because an overview of achieved specifications and errors will take place.

5.2.1 Quantitative Approach: Automatic MT Evaluation Metrics for ST-1’s Translation Outputs

This section will present the detailed results of MT outputs for BLEU and COMET metrics. Below is a table of the results of metrics -which are the quantitative evaluation tools used for this paper- for MT outputs against ST-1:

Table 13: Metric Evaluation Results of ST-1’s Candidate Texts

Output	1-gram (cumulative)	4-gram (cumulative)	BP	Overall BLEU	BLEU Rank	COMET-22	COMET Rank
Google Translate- raw	77.78	48.99	100.00	48.99	5	0.8521	8
GPT-4- zero-shot	68.57	29.91	100.00	29.91	9	0.8523	7
GPT-4- RBP	69.44	37.09	100.00	37.09	7	0.8725	5

Table 13 - continue

Gemini- 2.0-flash- zero-shot	76.42	56.03	100.00	56.03	2	0.8835	1
Gemini- 2.0-flash- RBP	75.73	51.41	100.00	51.41	4	0.8740	2
DeepSeek- R1-zero- shot	66.36	33.19	100.00	33.19	8	0.8457	9
DeepSeek- R1-RBP	67.27	38.78	100.00	38.78	6	0.8682	6
Claude- 3.5- Sonnet- zero-shot	78.10	57.14	100.00	57.14	1	0.8739	3
Claude-3.5- Sonnet- RBP	74.31	51.73	100.00	51.73	3	0.8732	4

Tilde MT's¹⁴ BLEU (Bilingual Evaluation Understudy) evaluator is used for automatic evaluation. BLEU is metric used to measure the quality of machine translations by comparing them to a reference (human) translation. It checks how similar the machine-generated translation is to the reference text based on overlapping words (n-grams) and phrases. The COMET-22 (wmt-comet-da5) metric was implemented in the Python environment of Google Colab to compute the evaluation results.¹⁵

¹⁴ <https://translate.tilde.ai/bleu#/>

¹⁵ <https://huggingface.co/Unbabel/wmt22-comet-da>

Interactive BLEU score evaluator

Perform comparative quality evaluations of files translated with one or more MT systems. This allows you to compare MT output with human translations and compare the BLEU scores of various MT systems. [Click here to learn more.](#)

Step 0: Pick source file (Optional)

Choose File no file selected

Step 1: Pick human translated file

Choose File ref

Step 2: Pick machine translated file

Choose File gemini-2.0-flash-raw

Step 3: Pick second machine translated file (Optional)

Choose File no file selected

Lowercase ☒

Tokenized ☒

Difference highlighting ☒

Calculate BLEU ☒

Display ☐

Score

Figure 6: Interface of BLEU Evaluator of Tilde MT with Uploaded Candidate and Reference Texts

One of the key components of BLEU metric is “n-gram”s. An n-gram is a sequence of n-consecutive words in a sentence. It helps BLEU compare short phrases between the machine translation and the reference text.

- **1-gram (Unigram):** single word comparison

Table 14: Visualization of 1-gram Matching in BLEU Metric Calculation

Example:	Reference (human) translation: Dakarbazin karşısında randomize faz 3 çalışma (CA209066) Candidate (machine) translation: Dakarbazine'ye karşı randomize faz 3 çalışma (CA209066)
----------	--

- **2-gram (Bigram):** two-word phrases (e.g., “randomize faz”, “faz 3”, “3 çalışma” “çalışma (CA209066)”)
- **3-gram (Trigram):** three-word phrases (e.g., “randomize faz 3”, “faz 3 çalışma”, “3 çalışma (CA209066)”)

- **4-gram:** four-word phrases (e.g., “randomize faz 3 çalışma”, “faz 3 çalışma (CA209066)”))

One disadvantageous aspect for agglutinative languages like Turkish, when it comes to n-gram evaluation is that, the more overlapping n-grams between the machine and reference translation, the higher the BLEU score – which can be misleading since agglutinative structures result in long, morphologically complex words that may not always align perfectly with the reference, even if the meaning is accurate. BLEU metric is about “exact” n-gram matches. Thus, this often leads to lower BLEU scores, not because of semantic errors, but due to structural mismatches caused by differences in word segmentation and morphology.

For instance, in the above 1-gram example, in the phrases “dakarbazin karşısında” and “dakarbazine’ye karşı” there is no n-gram match due to suffixational mismatches, although the collocations are semantically similar. This is a major disadvantage for agglutinative languages like Turkish. For this reason, some studies in the literature such as Ekinci’s work (2022:25) tend to lean towards 1-gram similarities rather than overall BLEU score because the overall tends to be lower as 4-gram similarity is low in such cases.

When we look at the above metrics, Google’s large language model Gemini-2.0-flash’s raw output is the 1st in BLEU among all the actuators. When we look at 1-gram results, however, one can see that Google Translate takes on the lead with 77.78 points of 1-gram score. It should be noted, however, that quantitative metrics do not always necessarily mean adequate translation has been achieved. Thus, in the upcoming sections, quantitative approach will be adopted as well as commentary section to see how well semantic similarity has been captured between reference text and candidate MT outputs.

5.2.2 Qualitative Approach: Customized Error-typology Framework (ST-1)

Table 15: TQE Overview of Penalty Scores of ST-1

Output	Total Penalty Points ↓	Custom TQE Rank ↑
Google Translate-raw	22	8

Table 15 - continue

GPT-4-zero-shot	19	6
GPT-4-RBP	21	7
Gemini-2.0-flash-zero-shot	10	4
Gemini-2.0-flash-RBP	12	5
DeepSeek-R1- zero-shot	12	5
DeepSeek-R1-RBP	6	3
Claude-3.5-Sonnet-zero-shot	0	1
Claude-3.5.-Sonnet-RBP	2	2

5.2.3 Quantitative Approach: Automatic MT Evaluation Metrics for ST-2's Translation Outputs

Table 16: Metric Evaluation Results of ST-2's Candidate Texts

Output	1-gram (cumulative)	4-gram (cumulative)	BP	Overall BLEU	BLEU Rank	COMET- 22¹⁶	COMET Rank
Google Translate- raw	56.63	24.46	100.00	24.46	7	NA	NA
GPT-4- zero-shot	60.47	33.01	100.00	33.01	4	NA	NA
GPT-4- RBP	63.86	38.93	100.00	38.93	1	NA	NA
Gemini- 2.0-flash- zero-shot	50.62	20.21	100.00	20.21	9	NA	NA

¹⁶ COMET scores of TT's for this ST were not calculated due to length of the text. As for automatic MT evaluation, only BLEU metric was used.

Table 16 - continue

Gemini-2.0-flash-RBP	54.32	21.36	100.00	21.36	8	NA	NA
DeepSeek-R1-zero-shot	62.07	34.89	100.00	34.89	2	NA	NA
DeepSeek-R1-RBP	57.30	32.75	100.00	32.75	5	NA	NA
Claude-3.5-Sonnet-zero-shot	59.76	28.88	100.00	28.88	6	NA	NA
Claude-3.5-Sonnet-RBP	59.09	33.69	100.00	33.69	3	NA	NA

5.2.4 Qualitative Approach: Customized Error-typology Framework (ST-2)

Table 17: TQE Overview of Penalty Scores of ST-2

Output	Total Penalty Points ↓	Custom TQE Rank ↑
Google Translate-raw	19	6
GPT-4-zero-shot	18	5
GPT-4-RBP	12	3
Gemini-2.0-flash-zero-shot	23	7
Gemini-2.0-flash-RBP	17	4
DeepSeek-R1-zero-shot	10	1
DeepSeek-R1-RBP	11	2
Claude-3.5-Sonnet-zero-shot	12	3

Table 17 - continue

Claude-3.5.-Sonnet-RBP	18	5
------------------------	----	---

5.2.5 Quantitative Approach: Automatic MT Evaluation Metrics for ST-3's Translation Outputs

Table 18: Metric Evaluation Results of ST-3's Candidate Texts

Output	1-gram (cumulative)	4-gram (cumulative)	BP	Overall BLEU	BLEU Rank	COMET-22	COMET Rank
Google Translate- raw	65.43	27.48	93.29	27.48	5	NA	NA
GPT-4- zero-shot	62.94	31.33	96.03	31.33	3	NA	NA
GPT-4- RBP	59.03	23.82	95.35	23.82	7	NA	NA
Gemini- 2.0-flash- zero-shot	65.58	20.43	100.00	20.43	9	NA	NA
Gemini- 2.0-flash- RBP	73.08	38.95	100.00	38.95	2	NA	NA
DeepSeek- R1-zero- shot	68.81	39.61	97.37	39.61	1	NA	NA
DeepSeek- R1-RBP	62.50	27.05	100.00	27.05	6	NA	NA

Table 18 - continue

Claude-3.5-Sonnet-zero-shot	62.87	29.55	93.98	29.55	4	NA	NA
Claude-3.5-Sonnet-RBP	58.92	23.39	92.60	23.39	8	NA	NA

5.2.6 Qualitative Approach: Customized Error-typology Framework (ST-3)

Table 19: TQE Overview of Penalty Scores of ST-3

Output	Total Penalty Points ↓	Custom TQE Rank ↑
Google Translate-raw	32	8
GPT-4- zero-shot	25	6
GPT-4-RBP	14	4
Gemini-2.0-flash-zero-shot	21	5
Gemini-2.0-flash-RBP	29	7
DeepSeek-R1-zero-shot	9	2
DeepSeek-R1-RBP	9	2
Claude-3.5-Sonnet-zero-shot	13	3
Claude-3.5.-Sonnet-RBP	6	1

5.3 Commentary on Findings

This section and its subsections are designed to elaborate on the MT outputs in a detailed fashion. Following subsections will analyze and comment on the translation decisions and translation errors in MT outputs of both ST's, with focus on both quantitative metrics and qualitative results derived from customized error-typology.

5.3.1 Analysis on ST-1's TT Outputs

This subsection will review 9 MT outputs against ST-1.

5.3.1.1 Google Translate – raw

Table 20: ST-1, Reference Text for ST-1, and MT Output (Google Translate – raw)

Segment No.	Source Text:	Reference Text:	MT Output:
1.	Randomised phase 3 study vs. dacarbazine (CA209066)	Dakarbazin karşısında randomize faz 3 çalışma (CA209066)	Randomize faz 3 çalışması ve dakarbazin (CA209066)
2.	The safety and efficacy of nivolumab 3 mg/kg for the treatment of advanced (unresectable or metastatic) melanoma were evaluated in a phase 3, randomised, double-blind study (CA209066).	Nivolumab 3 mg/kg'ın ilerlemiş (rezeke edilemeyen veya metastatik) melanom tedavisi için güvenliliği ve etkililiği bir faz 3, randomize, çift kör çalışmada (CA209066) değerlendirilmiştir.	İleri (çıkarılamaz veya metastatik) melanom tedavisinde nivolumab 3 mg/kg'ın güvenliği ve etkinliği faz 3, randomize, çift kör bir çalışmada (CA209066) değerlendirildi.
3.	The study included adult patients (18 years or older) with confirmed, treatment-naive, Stage III or IV BRAF wild-type melanoma and an Eastern Cooperative Oncology Group (ECOG) performance-status score of 0 or 1.	Çalışmaya doğrulanmış, tedavi uygulanmamış, Evre III veya IV BRAF vahşi tip melanomu olan ve ECOG performans durumu 0 veya 1 olan yetişkin hastalar (18 yaş ve üzeri) dahil edilmiştir.	Çalışmaya doğrulanmış, tedavi görmemiş, Evre III veya IV BRAF vahşi tip melanomu ve 0 veya 1 Doğu Kooperatif Onkoloji Grubu (ECOG) performans durumu puanı olan yetişkin hastalar (18 yaş ve üzeri) dahil edildi.

Table 20 - continue

4.	Patients with active autoimmune disease, ocular melanoma, or active brain or leptomeningeal	Aktif otoimmün hastalığı, oküler melanomu veya aktif beyin ya da leptomeningeal	Aktif otoimmün hastalığı, oküler melanomu veya aktif beyin veya leptomeningeal.
	metastases were excluded from the study.	metastazları bulunan hastalar çalışmaya alınmamıştır.	metastazları olan hastalar çalışmadan hariç tutuldu.

○ Error-typology chart of the MT output:

Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use				0	0
Terminology	Incorrect term use	2	2		4	8
Accuracy	Mistranslation		1		1	3
Accuracy	Overtranslation				0	0
Accuracy	Undertranslation				0	0
Accuracy	Addition				0	0
Accuracy	Omission				0	0
Accuracy	Untranslated				0	0
Linguistic Conver	Grammar	3			3	3
Linguistic Conver	Spelling				0	0
Syntax	Syntactical errors		1		1	3
Style & Clarity	Unnatural Phrasing	1	1		2	4
Style & Clarity	Poor Readability	1			1	1
Custom	Layout, character & other				0	0
		Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
		7	5	0	12	22

Figure 7: Error-typology Chart (Google Translate – raw)

Segment 1 is originally the sub-heading of the source passage from which the complete extract was taken. The preposition “vs.” denotes a comparison in the clinical trial in question (as can be inferred from the term “karşısında” in the reference text). Dacarbazine is the comparator drug and the preposition “vs.” signifies that the phase III trial is conducted against the comparator drug, whereas, the MT output of Google Translate renders the preposition as “ve” which is in fact “and” conjunction in English. This error changes the

intended meaning of the clinical trial description which can mislead the reader into thinking that dacarbazine is used alongside the primary treatment rather than being the comparator drug.

The collocation in the Segment 2 of the MT output: “ileri melanom tedavisinde” is also problematic due to the adjective “ileri”. In Turkish, “ileri” can indicate a positive progress or advancement, which may carry a positive connotation, potentially misleading the reader. The correct term should have been “ilerlemiş melanom”, which accurately reflects the clinical condition as advanced melanoma without the unintended positive meaning.

In the same segment again, the term “unresectable” was translated as “çıkarılamaz”, which is a simplified equivalent of the more accurate medical term “rezeke edilemeyen”. While “çıkarılamaz” conveys the general meaning, it lacks the precision expected in medical context. The term “rezeke edilemeyen” is standard and common ground among healthcare professionals. Using simplified language in such cases can lead to a loss of credibility of the medical text.

Safety and efficacy are two concepts that are very common in clinical trials. Their widely recognized counterparts are “güvenlilik” and “etkililik” in Turkish clinical trial settings. However, these terms are sometimes mistranslated as “güvenlik” and “etkinlik”. “Güvenlik” typically refers to physical security or protection, whereas “güvenlilik” is the correct term for the safety profile of a medical intervention. Similarly, “etkinlik” denotes activeness while “etkililik” specifically indicates the therapeutic efficacy of a treatment in achieving the intended clinical outcome. In Segment 2, these were translated as “güvenlik” and “etkinlik”.

In Segment 3, “Evre III veya IV BRAF vahşi tip melanomu ve 0 veya 1 Doğu Kooperatif Onkoloji Grubu (ECOG) performans durumu puanı olan yetişkin hastalar” has a complex syntax which makes readability difficult. Also in the same segment, ECOG (Eastern Cooperative Oncology Group) was translated word-for-word whereas there is no official Turkish counterpart for this organization.

In Segment 4, the sequence “Patients with active autoimmune disease, ocular melanoma, or active brain or leptomeningeal metastases” was rendered as “Aktif otoimmün

hastalığı, oküler melanomu veya aktif beyin veya leptomeningeal metastazları olan hastalar”. In the reference text one “veya” and one “ya da” were used for a better clarity in terms of grouping of the parts of speech. However, in the MT output, there are duplicate “veya” conjunctions which slightly hinders the readability. In the same segment, the verb “were excluded” should have been typically rendered as “çalışmadan dışlanmıştır”, or “çalışmaya alınmamıştır” which are common in clinical trial settings. The collocation “hariç tutuldu” in the MT output is not quite natural.

Lastly, the ST has a formal register which I did not have an option to (unless having trained NMT in a special way) communicate this (prompt, namely) to the NMT. All the past participles in verb endings in Turkish were translated as “-di”, rather than “-miştir/mıştır” which constitutes a formal register in such technical texts.

The above findings reveal the limitations of using Google Translate for such clinical trial extracts/texts. The errors identified highlight that the engine is not sufficiently equipped to handle domain-specific language with the required level of precision and formal register. Mistranslations and unnatural phrasing in such contexts can lead to misinterpretations, reducing the clarity and reliability of the output. Given the nature of the text, the MT output must undergo full post-editing.

Although the BLEU score was among the highest, the customized TQE analysis shows us the need for a more detailed error-focused evaluation. This analysis revealed that, despite the high BLEU score, the output’s reliability was questionable, highlighting the importance of quality assessment beyond automated metrics.

5.3.1.2 GPT-4 – zero-shot

Table 21: ST-1, Reference text for ST-1, and MT Output (GPT-4 – zero-shot)

Segment No.	Source Text:	Reference Text:	MT Output:

Table 21 - continue

1.	Randomised phase 3 study vs. dacarbazine (CA209066)	Dakarbazin karşısında randomize faz 3 çalışma (CA209066)	Faz 3 rastgele çalışma vs. dakarbazin (CA209066)
2.	The safety and efficacy of nivolumab 3 mg/kg for the treatment of advanced (unresectable or metastatic) melanoma were evaluated in a	Nivolumab 3 mg/kg'ın ilerlemiş (rezeke edilemeyen veya metastatik) melanom tedavisi için güvenliliği ve	İleri derecede (resekte edilemeyen veya metastatik) melanomun tedavisi için nivolumab 3 mg/kg'ın güvenliği ve etkinliği, faz 3,
	phase 3, randomised, double-blind study (CA209066).	etkililiği bir faz 3, randomize, çift kör çalışmada (CA209066) değerlendirilmiştir.	rastgele, çift-kör bir çalışma (CA209066) ile değerlendirilmiştir.
3.	The study included adult patients (18 years or older) with confirmed, treatment-naive, Stage III or IV BRAF wild-type melanoma and an Eastern Cooperative Oncology Group (ECOG) performance-status score of 0 or 1.	Çalışmaya doğrulanmış, tedavi uygulanmamış, Evre III veya IV BRAF vahşi tip melanomu olan ve ECOG performans durumu 0 veya 1 olan yetişkin hastalar (18 yaş ve üzeri) dahil edilmiştir.	Çalışma, teyit edilmiş, tedavi almamış, BRAF vahşi tip, Evre III veya IV melanomlu ve Doğu İşbirliği Onkoloji Grubu (ECOG) performans durumu skoru 0 veya 1 olan 18 yaş ve üzeri yetişkin hastaları içermiştir.
4.	Patients with active autoimmune disease, ocular melanoma, or active brain or leptomeningeal metastases were excluded from the study.	Aktif otoimmün hastalığı, oküler melanomu veya aktif beyin ya da leptomeningeal metastazları bulunan hastalar çalışmaya alınmamıştır.	Aktif otoimmün hastalığı, oküler melanomu veya aktif beyin veya leptomeningeal metastazı olan hastalar çalışmadan dışlanmıştır.

- Error-typology chart of the MT output:

Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use				0	0
Terminology	Incorrect term use	2	3		5	11
Accuracy	Mistranslation		1		1	3
Accuracy	Overtranslation				0	0
Accuracy	Undertranslation				0	0
Accuracy	Addition				0	0
Accuracy	Omission				0	0
Accuracy	Untranslated		1		1	3
Linguistic Converter	Grammar				0	0
Linguistic Converter	Spelling	1			1	1
Syntax	Syntactical errors				0	0
Style & Clarity	Unnatural Phrasing				0	0
Style & Clarity	Poor Readability	1			1	1
Custom	Layout, character & other				0	0
		Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
		4	5	0	9	19

Figure 8: Error-typology Chart (GPT-4 – zero-shot)

In Segment 1, it can be seen that the term “randomized” in the ST title was translated as “rastgele” which is a non-technical term for such context. In clinical trial terminology, the correct equivalent would be “randomize” which accurately reflects the methodology and controlled process of assigning patients to different treatment groups. Also in the segment, the preposition “vs.” was left untranslated.

The term “unresectable” was misspelled as “resekte edilemeyen” instead of “rezeke” or “rezekte” -which are both common-, which was classified as a minor spelling error. Same error as Google Translate’s output is observed: mistranslation as “güvenlik ve etkinlik” for “safety and efficacy”.

In Segment 3, the term “confirmed” is seen to be translated as “teyit edilmiş” instead of “doğrulanmış”. While “teyit edilmiş” conveys a sense of confirmation, it is more commonly used in everyday contexts and implies checking or verifying information directly with a person, such as a patient. However, in clinical discourse, “doğrulanmış” is the more precise term, as it refers to an objective validation process based on medical evidence, laboratory results, or clinical assessments rather than personal confirmation.

5.3.1.3 GPT-4 – RBP

Table 22: ST-1, Reference Text for ST-1, and MT Output (GPT-4 - RBP)

Segment No.	Source Text:	Reference Text:	MT Output:
1.	Randomised phase 3 study vs. dacarbazine (CA209066)	Dakarbazin karşısında randomize faz 3 çalışma (CA209066)	Randomize Faz 3 Çalışması vs. Dakarbazin (CA209066)
2.	The safety and efficacy of nivolumab 3 mg/kg for the treatment of advanced (unresectable or metastatic) melanoma were evaluated in a phase 3, randomised, double-blind study (CA209066).	Nivolumab 3 mg/kg'ın ilerlemiş (rezeke edilemeyen veya metastatik) melanom tedavisi için güvenliliği ve etkililiği bir faz 3, randomize, çift kör çalışmada (CA209066) değerlendirilmiştir.	İleri (rezeksiyon yapılamayan veya metastatik) melanomun tedavisi için nivolumab 3 mg/kg'ın güvenilirliği ve etkinliği, faz 3, randomize, çift-kör çalışma (CA209066) ile değerlendirilmiştir.
3.	The study included adult patients (18 years or older) with confirmed, treatment-naive, Stage III or IV BRAF wild-type melanoma and an Eastern Cooperative Oncology Group (ECOG) performance-status score of 0 or 1.	Çalışmaya doğrulanmış, tedavi uygulanmamış, Evre III veya IV BRAF vahşi tip melanomu olan ve ECOG performans durumu 0 veya 1 olan yetişkin hastalar (18 yaş ve üzeri) dahil edilmiştir.	Çalışma, onaylanmış, tedavi-naif, Evre III veya IV BRAF vahşi tip melanomu olan ve Doğu İşbirliği Onkoloji Grubu (ECOG) performans durumu skoru 0 veya 1 olan yetişkin hastaları (18 yaş ve üzeri) kapsamıştır.
4.	Patients with active autoimmune disease, ocular melanoma, or active brain or leptomeningeal metastases were excluded from the study.	Aktif otoimmün hastalığı, oküler melanomu veya aktif beyin ya da leptomeningeal metastazları bulunan	Aktif otoimmün hastalığı olan, göz melanomu olan veya aktif beyin veya leptomeningeal metastazları olan hastalar çalışmaya dahil edilmemiştir.

Table 22 - continue

		hastalar çalışmaya alınmamıştır.	
--	--	----------------------------------	--

- Error-typology chart of the MT output:

Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use				0	0
Terminology	Incorrect term use	1	2		3	7
Accuracy	Mistranslation		1		1	3
Accuracy	Overtranslation				0	0
Accuracy	Undertranslation				0	0
Accuracy	Addition				0	0
Accuracy	Omission				0	0
Accuracy	Untranslated		1		1	3
Linguistic Converter	Grammar				0	0
Linguistic Converter	Spelling				0	0
Syntax	Syntactical errors				0	0
Style & Clarity	Unnatural Phrasing	1	1		2	4
Style & Clarity	Poor Readability	1	1		2	4
Custom	Layout, character & other				0	0
		Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
		3	6	0	9	21

Figure 9: Error-typology Chart (GPT-4 – RBP)

Segment 1 puts forth appropriate terminology use with the translation of “randomized”, however, the preposition “vs.” is left untranslated in the MT output, leading to a loss of comparison meaning. In Segment 2, the phrase “rezeksiyon yapılamayan” is considered an acceptable rendering, as “rezeksiyon” accurately denotes the surgical procedure of “rezeke etmek”. A notable error in this segment is the translation of “güvenilirliği”, which, in Turkish, corresponds to “reliability” or “credibility” rather than the intended meaning of “safety” within the clinical context.

In Segment 3, the terms “onaylanmış” and “tedavi-naif” are problematic inaccuracies. The use of “onaylanmış” is misleading, as it conveys the notion of something being “accepted” or “approved” rather than “validated” through scientific evidence. Similarly, “tedavi-naif” constitutes an unnatural and non-standard use in Turkish language. This compound word fails to align with the linguistic conventions of the target language, and a

more appropriate alternative would be “tedavi almamış”, which accurately conveys the concept of a treatment-naive patient in a manner consistent with Turkish medical discourse. The mistranslation error for the acronym “ECOG” still persists like the previous outputs. Lastly, the same error with the use of duplicate conjunction “veya” is also observed in this output.

5.3.1.4 Gemini-2.0-flash – zero-shot

Table 23: ST-1, Reference Text for ST-1, and MT Output (Gemini-2.0-Flash – zero-shot)

Segment No.	Source Text:	Reference Text:	MT Output:
1.	Randomised phase 3 study vs. dacarbazine (CA209066)	Dakarbazin karşısında randomize faz 3 çalışma (CA209066)	Dakarbazine’ye karşı randomize faz 3 çalışma (CA209066)
2.	The safety and efficacy of nivolumab 3 mg/kg for the treatment of advanced (unresectable or metastatic) melanoma were evaluated in a phase 3, randomised, double-blind study (CA209066).	Nivolumab 3 mg/kg’in ilerlemiş (rezeke edilemeyen veya metastatik) melanom tedavisi için güvenliliği ve etkililiği bir faz 3, randomize, çift kör çalışmada (CA209066) değerlendirilmiştir.	İleri (rezeke edilemeyen veya metastatik) melanom tedavisi için 3 mg/kg nivolumab’ın güvenliği ve etkinliği, faz 3, randomize, çift kör bir çalışmada (CA209066) değerlendirilmiştir.
3.	The study included adult patients (18 years or older) with confirmed, treatment-naive, Stage III or IV BRAF wild-type melanoma and an Eastern Cooperative Oncology Group	Çalışmaya doğrulanmış, tedavi uygulanmamış, Evre III veya IV BRAF vahşi tip melanomu olan ve ECOG performans durumu 0 veya 1 olan yetişkin hastalar (18	Çalışmaya, doğrulanmış, tedavi görmemiş, Evre III veya IV BRAF vahşi tip melanomu olan ve Doğu Kooperatif Onkoloji Grubu (ECOG) performans durumu

Table 23 - continue

	(ECOG) performance-status score of 0 or 1.	yaş ve üzeri) dahil edilmiştir.	skoru 0 veya 1 olan yetişkin hastalar (18 yaş veya üstü) dahil edilmiştir.
4.	Patients with active autoimmune disease, ocular melanoma, or active brain or leptomeningeal metastases were excluded from the study.	Aktif otoimmün hastalığı, oküler melanomu veya aktif beyin ya da leptomeningeal metastazları bulunan hastalar çalışmaya alınmamıştır.	Aktif otoimmün hastalığı, oküler melanomu veya aktif beyin veya leptomeningeal metastazları olan hastalar çalışma dışı bırakılmıştır.

○ Error-typology chart of the MT output:

Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use				0	0
Terminology	Incorrect term use	2			2	2
Accuracy	Mistranslation		1		1	3
Accuracy	Overtranslation				0	0
Accuracy	Undertranslation				0	0
Accuracy	Addition				0	0
Accuracy	Omission				0	0
Accuracy	Untranslated				0	0
Linguistic Converter	Grammar				0	0
Linguistic Converter	Spelling	1			1	1
Syntax	Syntactical errors				0	0
Style & Clarity	Unnatural Phrasing		1		1	3
Style & Clarity	Poor Readability	1			1	1
Custom	Layout, character & other				0	0
		Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
		4	2	0	6	10

Figure 10: Error-typology Chart (Gemini-2.0-flash - zero-shot)

The title in Segment 1 shows that the proper noun for the compound “dacarbazine” has an orthographic problem when rendered into Turkish. It is seen to be rendered as “dakarbazine” as one can infer from the following phrase “Dakarbazine’ye”. “Dakarbazin” is the standard Turkish form of the compound and it does not act as a proper noun

gramatically, therefore, it doesn't need an apostrophe for a suffix following it as in "Dakarbazine'ye". Instead of this, it should have been written as "dakarbazine karşı" (vs. dacarbazine).

Segment 3 is notably successful in translation of the following concepts: confirmed and treatment-naïve as in the phrase: "Çalışmaya doğrulanmış, tedavi görmemiş (...)" while the errors "Doğu Kooperatif Onkoloji Grubu (ECOG)" and duplicate use of "veya" conjunction persist.

5.3.1.5 Gemini-2.0-flash – RBP

Table 24: ST-1, Reference Text for ST-1, and MT Output (Gemini-2.0-flash - RBP)

Segment No.	Source Text:	Reference Text:	MT Output:
1.	Randomised phase 3 study vs. dacarbazine (CA209066)	Dakarbazin karşısında randomize faz 3 çalışma (CA209066)	Randomize faz 3 çalışması vs. dakarbazin (CA209066)
2.	The safety and efficacy of nivolumab 3 mg/kg for the treatment of advanced (unresectable or metastatic) melanoma were evaluated in a phase 3, randomised, double-blind study (CA209066).	Nivolumab 3 mg/kg'ın ilerlemiş (rezeke edilemeyen veya metastatik) melanom tedavisi için güvenliliği ve etkililiği bir faz 3, randomize, çift kör çalışmada (CA209066) değerlendirilmiştir.	İleri (rezeke edilemeyen veya metastatik) melanom tedavisinde nivolumab 3 mg/kg'ın güvenliliği ve etkinliği, randomize, çift kör, faz 3 çalışmasında (CA209066) değerlendirilmiştir.
3.	The study included adult patients (18 years or older) with confirmed, treatment-naïve, Stage III or IV BRAF wild-type melanoma and an Eastern	Çalışmaya doğrulanmış, tedavi uygulanmamış, Evre III veya IV BRAF vahşi tip melanomu olan ve ECOG performans durumu 0 veya	Çalışmaya, doğrulanmış, tedavi almamış, Evre III veya IV BRAF vahşi tip melanomu olan ve Doğu Kooperatif Onkoloji Grubu (ECOG)

Table 24 - continue

	Cooperative Oncology Group (ECOG) performance-status score of 0 or 1.	1 olan yetişkin hastalar (18 yaş ve üzeri) dahil edilmiştir.	performans durumu skoru 0 veya 1 olan yetişkin hastalar (18 yaş veya üstü) dahil edilmiştir.
4.	Patients with active autoimmune disease, ocular melanoma, or active brain or leptomeningeal metastases were excluded from the study.	Aktif otoimmün hastalığı, oküler melanomu veya aktif beyin ya da leptomeningeal metastazları bulunan hastalar çalışmaya alınmamıştır.	Aktif otoimmün hastalığı, oküler melanomu veya aktif beyin veya leptomeningeal metastazları olan hastalar çalışma dışı bırakılmıştır.

- Error-typology chart of the MT output:

Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use				0	0
Terminology	Incorrect term use	1			1	1
Accuracy	Mistranslation		1		1	3
Accuracy	Overtranslation				0	0
Accuracy	Undertranslation				0	0
Accuracy	Addition				0	0
Accuracy	Omission				0	0
Accuracy	Untranslated		1		1	3
Linguistic Converter	Grammar				0	0
Linguistic Converter	Spelling				0	0
Syntax	Syntactical errors				0	0
Style & Clarity	Unnatural Phrasing	1	1		2	4
Style & Clarity	Poor Readability	1			1	1
Custom	Layout, character & other				0	0
		Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
		3	3	0	6	12

Figure 11: Error-typology Chart (Gemini-2.0-flash – RBP)

Overall MT output in this one is quite similar to the result obtained in Gemini-2.0-flash – raw output. In Segment 1, it can be seen that the preposition “vs.” remains untranslated. The problematic collocation in Segment 2: “ileri melanom” still persists as in

the majority of the outputs. Safety and efficacy was translated in the raw Gemini output as “güvenliği ve etkinliği” whereas the role-based output reads as “güvenliliği ve etkinliği”, giving the accurate counterpart for safety, yet, efficacy is still inaccurate. The majority of the errors between raw and role-based outputs from Gemini are the same.

5.3.1.6 DeepSeek – R1 – zero-shot

Table 25: ST-1, Reference Text for ST-1, and MT Output (DeepSeek-R1 – zero-shot)

Segment No.	Source Text:	Reference Text:	MT Output:
1.	Randomised phase 3 study vs. dacarbazine (CA209066)	Dakarbazin karşısında randomize faz 3 çalışma (CA209066)	Randomize faz 3 çalışma: Dakarbazin ile karşılaştırmalı (CA209066)
2.	The safety and efficacy of nivolumab 3 mg/kg for the treatment of advanced (unresectable or metastatic) melanoma were evaluated in a phase 3, randomised, double-blind study (CA209066).	Nivolumab 3 mg/kg’ın ilerlemiş (rezeke edilemeyen veya metastatik) melanom tedavisi için güvenliliği ve etkililiği bir faz 3, randomize, çift kör çalışmada (CA209066) değerlendirilmiştir.	İleri evre (cerrahi olarak çıkarılamayan veya metastatik) melanom tedavisi için nivolumab’ın 3 mg/kg dozunun güvenliliği ve etkinliği, faz 3, randomize, çift kör bir çalışmada (CA209066) değerlendirilmiştir.
3.	The study included adult patients (18 years or older) with confirmed, treatment-naive, Stage III or IV BRAF wild-type melanoma and an Eastern Cooperative Oncology Group (ECOG) performance-status score of 0 or 1.	Çalışmaya doğrulanmış, tedavi uygulanmamış, Evre III veya IV BRAF vahşi tip melanomu olan ve ECOG performans durumu 0 veya 1 olan yetişkin hastalar (18 yaş ve üzeri) dahil edilmiştir.	Çalışmaya, BRAF wild tip Evre III veya IV melanom teşhisi konulmuş, daha önce tedavi görmemiş, 18 yaş ve üzeri yetişkin hastalar ile Eastern Cooperative Oncology Group (ECOG) performans durum skoru 0

Table 25 - continue

			veya 1 olan bireyler dahil edilmiştir.
4.	Patients with active autoimmune disease, ocular melanoma, or active brain or leptomeningeal metastases were excluded from the study.	Aktif otoimmün hastalığı, oküler melanomu veya aktif beyin ya da leptomeningeal metastazları bulunan hastalar çalışmaya alınmamıştır.	Aktif otoimmün hastalığı bulunanlar, oküler melanom hastaları veya aktif beyin/leptomeningeal metastazı olanlar çalışma dışı bırakılmıştır.

- Error-typology chart of the MT output:

Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use				0	0
Terminology	Incorrect term use	1			1	1
Accuracy	Mistranslation				0	0
Accuracy	Overtranslation		1		1	3
Accuracy	Undertranslation		1		1	3
Accuracy	Addition				0	0
Accuracy	Omission				0	0
Accuracy	Untranslated				0	0
Linguistic Conve	Grammar				0	0
Linguistic Conve	Spelling				0	0
Syntax	Syntactical errors				0	0
Style & Clarity	Unnatural Phrasing				0	0
Style & Clarity	Poor Readability		1		1	3
Custom	Layout, character & other	1			1	1
		Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
		2	3	0	5	11

Figure 12: Error-typology chart (DeepSeek-R1 – zero-shot)

In the first segment, on the The MT output “Randomize faz 3 çalışma: Dakarbazin ile karşılaştırmalı (CA209066)” shows deviations from the reference text that affect clarity in the context. The MT output restructures the sentence, converting the original “vs.” (karşısında) into “ile karşılaştırmalı” and adding a colon after “faz 3 çalışma.” While “karşılaştırmalı” conveys comparison, this restructuring changes the directness of the

comparison, making it sound more like a descriptive phrase rather than indicating a controlled clinical trial against a comparator drug, which is dacarbazine. In addition to that, the reference text “karşısında” correctly captures the meaning of “vs.” by clearly positioning dacarbazine as the comparator drug in the trial. However, the MT output “ile karşılaştırmalı” shifts the emphasis and could be misinterpreted as a general comparative trial rather than a randomized controlled trial where dacarbazine acts as the comparator. Besides, “ile karşılaştırmalı” might suggest a more general comparison.

In Segment 2, the collocation “ileri evre melanom” provides a clearer output, compared to those which read “ileri melanom”, in the sense that it won’t lead to ambiguities in the classification of the disease “advanced melanoma”. Segment 3 of MT output reads as: “Çalışmaya, BRAF wild tip Evre III veya IV melanom teşhisi konulmuş, daha önce tedavi görmemiş, 18 yaş ve üzeri yetişkin hastalar ile Eastern Cooperative Oncology Group (ECOG) performans durum skoru 0 veya 1 olan bireyler dahil edilmiştir”. “Melanom teşhisi konulmuş” can be deemed a minor mistranslation as “confirmed” refers to the fact that the melanoma diagnosis on the participants included in the clinical trial underwent validation following the diagnosis at a later stage. However, “teşhisi konulmuş”, won’t necessarily indicate that the diagnosis of the disease underwent a later clinical assessment for validation or confirmation. Although moderate, this is a case of undertranslation or mistranslation.

The equivalent “tedavi görmemiş” for “treatment-naïve” is a correct collocation. Contrary to all the rest of the outputs in this experimental study, DeepSeek-R1’s raw carried out an accurate method for translating ECOG as in the following: “Eastern Cooperative Oncology Group (ECOG) performans durum skoru (...)” by first putting the complete name of the acronym and then giving the abbreviation of it in parentheses.

In addition, Segment 4 does not contain the duplicate use of “veya + veya”, common among other outputs, as in the following: “Aktif otoimmün hastalığı bulunanlar, oküler melanom hastaları veya aktif beyin/leptomeningeal metastazı olanlar çalışma dışı bırakılmıştır” which enables a clear readability rather than leading to poor readability.

5.3.1.7 DeepSeek – R1 – RBP

Table 26: ST-1, Reference Text for ST-1, and MT Output (DeepSeek-R1 - RBP)

Segment No.	Source Text:	Reference Text:	MT Output:
1.	Randomised phase 3 study vs. dacarbazine (CA209066)	Dakarbazin karşısında randomize faz 3 çalışma (CA209066)	Randomize Faz 3 Çalışma (Dakarbazin ile Karşılaştırma – CA209066)
2.	The safety and efficacy of nivolumab 3 mg/kg for the treatment of advanced (unresectable or metastatic) melanoma were evaluated in a phase 3, randomised, double-blind study (CA209066).	Nivolumab 3 mg/kg'ın ilerlemiş (rezeke edilemeyen veya metastatik) melanom tedavisi için güvenliliği ve etkililiği bir faz 3, randomize, çift kör çalışmada (CA209066) değerlendirilmiştir.	İleri evre (rezeke edilemeyen veya metastatik) melanom tedavisinde nivolumab'ın (3 mg/kg) güvenlilik ve etkinliği, CA209066 numaralı Faz 3, randomize, çift kör bir çalışmada değerlendirilmiştir.
3.	The study included adult patients (18 years or older) with confirmed, treatment-naive, Stage III or IV BRAF wild-type melanoma and an Eastern Cooperative Oncology Group (ECOG) performance-status score of 0 or 1.	Çalışmaya doğrulanmış, tedavi uygulanmamış, Evre III veya IV BRAF vahşi tip melanomu olan ve ECOG performans durumu 0 veya 1 olan yetişkin hastalar (18 yaş ve üzeri) dahil edilmiştir.	Çalışmaya, BRAF wild-type Evre III veya IV melanom tanısı almış, daha önce tedavi görmemiş, Doğu İşbirliği Onkoloji Grubu (ECOG) performans durum skoru 0 veya 1 olan yetişkin hastalar (18 yaş ve üzeri) dahil edilmiştir.
4.	Patients with active autoimmune disease, ocular melanoma, or active brain or leptomeningeal metastases were excluded from the study.	Aktif otoimmün hastalığı, oküler melanomu veya aktif beyin ya da leptomeningeal metastazları bulunan	Aktif otoimmün hastalığı bulunanlar, oküler melanomu olanlar veya aktif beyin/leptomeningeal

Table 26 - continue

		hastalar çalışmaya alınmamıştır.	metastazı olan hastalar çalışma dışı bırakılmıştır.
--	--	----------------------------------	---

- Error-typology chart of the MT output:

Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use				0	0
Terminology	Incorrect term use	1			1	1
Accuracy	Mistranslation		1		1	3
Accuracy	Overtranslation				0	0
Accuracy	Undertranslation				0	0
Accuracy	Addition				0	0
Accuracy	Omission				0	0
Accuracy	Untranslated	1			1	1
Linguistic Conver	Grammar				0	0
Linguistic Conver	Spelling				0	0
Syntax	Syntactical errors				0	0
Style & Clarity	Unnatural Phrasing				0	0
Style & Clarity	Poor Readability	1			1	1
Custom	Layout, character & other				0	0
		Minor Total	Major Total	Critical Tota	Total Error Instances	Final Penalty Score
		3	1	0	4	6

Figure 13: Error-typology Chart (DeepSeek-R1 - RBP)

One of the outputs with the least errors, RBP output from R1 model of DeepSeek has achieved a quite well-structured translation. In Segment 1, the title's core idea seems to be restrained in parantheses which sounds as though it was an additional information: "Randomize Faz 3 Çalışma (Dakarbazin ile Karşılaştırma – CA209066)", whereas dacarbazine being the comparator compound is one of the key ideas of the title. In Segment 2, "İleri evre melanom" serves as an appropriate equivalent for "advanced melanoma".

In Segment 3, the phrase "... wild-type Evre III veya IV melanom tanısı almış..." contains an untranslated collocation, namely "wild-type". While "wild-tip kanser" or "wild-tip BRAF melanom" are established terms in Turkish medical language and especially oncology sub-domain, "type" itself remains untranslated here. Even though this text is from an SmPC document -intended for healthcare professionals- this still constitutes an error. Even

if the intended audience can infer that it refers to “vahşi tip melanom” or “wild-tip melanom”, the term “wild-type” has not been directly coined in Turkish medical literature, at least in written settings. If the phrase had been “wild-tip”, as seen in some other outputs in this study, it would have aligned with established usage.

Unlike many other outputs, the acronym “ECOG” in Segment 3 is correctly left as is. Additionally, in Segment 4, the duplicate “veya” usage has been resolved. Overall, this output seems to have eliminated the majority of the common issues in other MT outputs.

5.3.1.8 Claude-3.5-Sonnet – zero-shot

Table 27: ST-1, Reference Text for ST-1, and MT Output (Claude-3.5-Sonnet – zero-shot)

Segment No.	Source Text:	Reference Text:	MT Output:
1.	Randomised phase 3 study vs. dacarbazine (CA209066)	Dakarbazin karşısında randomize faz 3 çalışma (CA209066)	Dakarbazine karşı randomize faz 3 çalışması (CA209066)
2.	The safety and efficacy of nivolumab 3 mg/kg for the treatment of advanced (unresectable or metastatic) melanoma were evaluated in a phase 3, randomised, double-blind study (CA209066).	Nivolumab 3 mg/kg’ın ilerlemiş (rezeke edilemeyen veya metastatik) melanom tedavisi için güvenliliği ve etkililiği bir faz 3, randomize, çift kör çalışmada (CA209066) değerlendirilmiştir.	İlerlemiş (rezeke edilemeyen veya metastatik) melanom tedavisinde nivolumab 3 mg/kg’ın güvenliliği ve etkililiği, faz 3, randomize, çift kör bir çalışmada (CA209066) değerlendirilmiştir.
3.	The study included adult patients (18 years or older) with confirmed, treatment-naive, Stage III or IV BRAF wild-type melanoma and an Eastern	Çalışmaya doğrulanmış, tedavi uygulanmamış, Evre III veya IV BRAF vahşi tip melanomu olan ve ECOG performans durumu 0 veya	Çalışmaya, doğrulanmış, tedavi görmemiş, Evre III veya IV BRAF wild-tip melanomu olan ve Eastern Cooperative Oncology Group

Table 27 - continue

	Cooperative Oncology Group (ECOG) performance-status score of 0 or 1.	1 olan yetişkin hastalar (18 yaş ve üzeri) dahil edilmiştir.	(ECOG) performans durumu skoru 0 veya 1 olan yetişkin hastalar (18 yaş ve üzeri) dahil edilmiştir.
4.	Patients with active autoimmune disease, ocular melanoma, or active brain or leptomeningeal metastases were excluded from the study.	Aktif otoimmün hastalığı, oküler melanomu veya aktif beyin ya da leptomeningeal metastazları bulunan hastalar çalışmaya alınmamıştır.	Aktif otoimmün hastalığı, oküler melanomu veya aktif beyin ya da leptomeningeal metastazı olan hastalar çalışma dışı bırakılmıştır.

- Error-typology chart of the MT output:

Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use				0	0
Terminology	Incorrect term use				0	0
Accuracy	Mistranslation				0	0
Accuracy	Overtranslation				0	0
Accuracy	Undertranslation				0	0
Accuracy	Addition				0	0
Accuracy	Omission				0	0
Accuracy	Untranslated				0	0
Linguistic Converter	Grammar				0	0
Linguistic Converter	Spelling				0	0
Syntax	Syntactical errors				0	0
Style & Clarity	Unnatural Phrasing				0	0
Style & Clarity	Poor Readability				0	0
Custom	Layout, character & other				0	0
		Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
		0	0	0	0	0

Figure 14: Error-typology Chart (Claude-3.5-Sonnet – zero-shot)

This output has been the only output with no errors detected among all the MT outputs. Title is accurately rendered. In Segment 2, “ilerlemiş melanom”, “güvenlilik” and “etkililik” are accurate equivalences which have been one of the most common incorrect term

usage errors of all the outputs.

As also mentioned in the previous commentary on DeepSeek-R1-RBP output, “wild-tip” is not deemed an incorrect usage because it is an established usage in medical literature within oncology sub-domain. If it had been translated, for instance, as: “wild-type” (such as in DeepSeek’s output) it would have been counted as untranslated due to the term “type”. That is to say, “wild-tip” is a common ground term established in medical literature, which implies a fast growing, highly invasive type of condition (such as melanoma or cancer in general).

In Segment 3, the acronym ECOG is left as it is. In Segment 4, the issue with duplicate use of “veya” has been overcome by using “veya” and “ya da” as in the following: “Aktif otoimmün hastalığı, oküler melanomu veya aktif beyin ya da leptomeningeal metastazı olan hastalar çalışma dışı bırakılmıştır.” Overall, the output has no translation errors detected.

5.3.1.9 Claude-3.5-Sonnet– RBP

Table 28: ST-1, Reference Text for ST-1, and MT Output (Claude-3.5-Sonnet - RBP)

Segment No.	Source Text:	Reference Text:	MT Output:
1.	Randomised phase 3 study vs. dacarbazine (CA209066)	Dakarbazin karşısında randomize faz 3 çalışma (CA209066)	Dakarbazin ile karşılaştırmalı randomize faz 3 çalışması (CA209066)
2.	The safety and efficacy of nivolumab 3 mg/kg for the treatment of advanced (unresectable or metastatic) melanoma were evaluated in a phase 3, randomised, double-blind study (CA209066).	Nivolumab 3 mg/kg’ın ilerlemiş (rezeke edilemeyen veya metastatik) melanom tedavisi için güvenliliği ve etkililiği bir faz 3, randomize, çift kör çalışmada (CA209066) değerlendirilmiştir.	İleri evre (rezeke edilemeyen veya metastatik) melanom tedavisinde nivolumab 3 mg/kg’ın güvenliliği ve etkililiği, faz 3, randomize, çift kör bir çalışmada (CA209066) değerlendirilmiştir.

Table 28 - continue

3.	The study included adult patients (18 years or older) with confirmed, treatment-naive, Stage III or IV BRAF wild-type melanoma and an Eastern Cooperative Oncology Group (ECOG) performance-status score of 0 or 1.	Çalışmaya doğrulanmış, tedavi uygulanmamış, Evre III veya IV BRAF vahşi tip melanomu olan ve ECOG performans durumu 0 veya 1 olan yetişkin hastalar (18 yaş ve üzeri) dahil edilmiştir.	Çalışmaya, doğrulanmış, daha önce tedavi almamış, Evre III veya IV BRAF wild-tip melanomu olan ve Eastern Cooperative Oncology Group (ECOG) performans durumu skoru 0 veya 1 olan erişkin hastalar (18 yaş ve üzeri) dahil edilmiştir.
4.	Patients with active autoimmune disease, ocular melanoma, or active brain or leptomeningeal metastases were excluded from the study.	Aktif otoimmün hastalığı, oküler melanomu veya aktif beyin ya da leptomeningeal metastazları bulunan hastalar çalışmaya alınmamıştır.	Aktif otoimmün hastalığı, oküler melanomu veya aktif beyin ya da leptomeningeal metastazı olan hastalar çalışma dışı bırakılmıştır.

Error-typology chart of the MT output:

Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use				0	0
Terminology	Incorrect term use				0	0
Accuracy	Mistranslation				0	0
Accuracy	Overtranslation				0	0
Accuracy	Undertranslation				0	0
Accuracy	Addition				0	0
Accuracy	Omission				0	0
Accuracy	Untranslated				0	0
Linguistic Conve	Grammar				0	0
Linguistic Conve	Spelling	1			1	1
Syntax	Syntactical errors				0	0
Style & Clarity	Unnatural Phrasing				0	0
Style & Clarity	Poor Readability	1			1	1
Custom	Layout, character & other				0	0
		Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
		2	0	0	2	2

Figure 15: Error-typology Chart (Claude-3.5-Sonnet - RBP)

Similar to the raw output, this role-based output from Claude’s Sonnet model has achieved a quality translation with almost no errors. There are only 2 errors observed: interestingly, when prompted as role-based, meaning prompted to take into account some considerations, the model translated the active substance with a slight misspelling as “dakarbasin”. The Turkicized version is misspelled as it should have been spelled as “dakarbazin”.

And similar to DeepSeek’s raw output from R1 model, the title in Segment 1 is translated as: “Dakarbasin ile karşılaştırmalı randomize faz 3 çalışması (CA209066)”. The phrase “ile karşılaştırmalı” leads to deviations from the reference text that impact clarity in the context, comparing to use of “karşısında” or “-e karşı” which convey the meaning “vs.”.

5.3.2 Overall Results Overview for ST-1’s TT Outputs

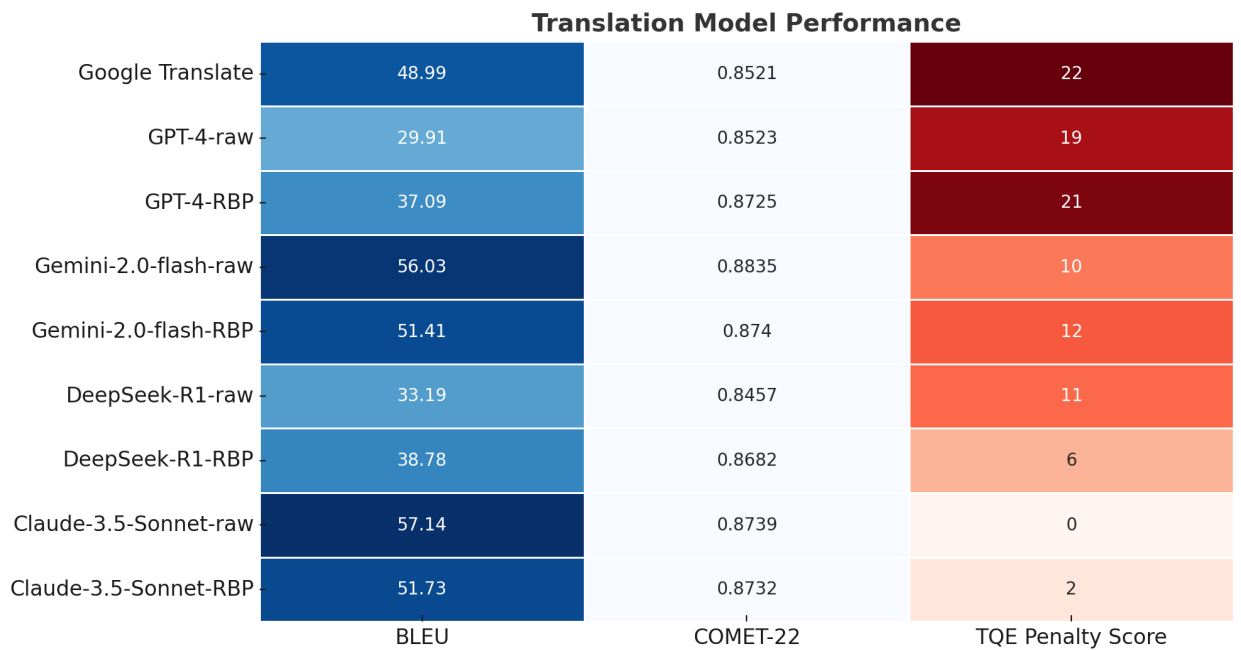


Figure 16: Heatmap Overview of Quantitative and Qualitative Metric Results of ST-1’s TT Outputs

5.3.3 Analysis on ST-2's TT Outputs

This subsection will review 9 MT outputs against ST-2. Due to the fact that there are some repetitive errors among the majority of the outputs, that fall into same error types and are abundant errors, I will not elaborate in a detailed manner on each individual output however will demonstrate fundamental errors and offer rationale for various comments.

5.3.3.1 Google Translate – raw

Table 29: ST-2, Reference Text for ST-2, and MT Output (Google Translate - raw)

Segment No.	Source Text:	Reference Text:	MT Output:
1.	Breast-feeding	Emzirme	Emzirme
2.	Do not breast-feed while you are taking this medicine.	Bu ilacı kullanırken emzirmeyiniz.	Bu ilacı kullanırken emzirmeyin.
3.	It is not known whether the active substance in Venclyxto passes into breast milk.	VENCLYXTO'nun etkin maddesinin anne sütüne geçip geçmediği bilinmemektedir.	Venclyxto'daki etkin maddenin anne sütüne geçip geçmediği bilinmemektedir.
4.	Fertility	Fertilite (üreme yeteneği)	Doğurganlık
5.	Based on findings in animals, Venclyxto may cause male infertility (low or no sperm count).	Hayvan çalışmaları bulgularına göre, VENCLYXTO erkek infertilitesine (sperm sayısında düşme ya da sperm yokluğu) neden olabilir.	Hayvanlarda yapılan bulgulara göre, Venclyxto erkek kısırlığına (düşük veya hiç sperm sayısı) neden olabilir.
6.	This may affect your ability to father a child.	Bu durum baba olma yetinizi etkileyebilir.	Bu, çocuk sahibi olma yeteneğinizi etkileyebilir.

Table 29 - continue

7.	Ask your doctor for advice on sperm storage before starting treatment with Venclyxto.	VENCLYXTO tedavisine başlamadan önce sperm depolama hakkında doktorunuza danışınız.	Venclyxto ile tedaviye başlamadan önce doktorunuzdan sperm depolaması hakkında tavsiye isteyin.
8.	Driving and using machines	Araç ve makine kullanımı	Araba ve makine kullanımı
9.	You may feel tired or dizzy after taking Venclyxto, which may affect your ability to drive or use tools or machines.	VENCLYXTO kullandıktan sonra, araç ve makine kullanımı kabiliyetinizi etkileyebilecek yorgunluk veya baş dönmesi hissedebilirsiniz.	Venclyxto aldıktan sonra yorgun veya baş dönmesi hissedebilirsiniz, bu da araba kullanma veya alet veya makine kullanma yeteneğinizi etkileyebilir.

○ Error-typology chart of the MT output:

Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use				0	0
Terminology	Incorrect term use	1	1		2	4
Accuracy	Mistranslation				0	0
Accuracy	Overtranslation				0	0
Accuracy	Undertranslation	1		1	2	6
Accuracy	Addition				0	0
Accuracy	Omission				0	0
Accuracy	Untranslated				0	0
Linguistic Conventions	Grammar	3			3	3
Linguistic Conventions	Spelling				0	0
Syntax	Syntactical errors				0	0
Style & Clarity	Unnatural Phrasing	2	1		3	5
Style & Clarity	Poor Readability	1			1	1
Custom	Layout, character & other				0	0
		Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
		8	2	1	11	19

Figure 17: Error-typology Chart (Google Translate - raw)

This output has a quite high TQE penalty score of 19. In Segments 2 and 7, the imperative verbs “emzirmeyin” (do not breast-feed) and “tavsiye isteyin” (ask for advice) end with the suffixes “-in” whereas for such documents as PILs, the imperative form “-iniz” would be preferred to maintain a more formal register. Although a minor modality error, this should be taken into account when writing a technical text as such.

In Segment 4, the sub-heading “Fertility” is translated in MT output as “Doğurganlık” which signifies reproduction of women through pregnancy, whereas the following section mentions the fact that male fertility can be affected by this medicine, along with the ability to father a child, based on findings in studies conducted on animals. Hence, this was a critical error leading to a crucial ambiguity since the section elaborates on male fertility whereas the term “Doğurganlık” is peculiar to female creatures. In the reference segment, one can see that it is rendered as “Fertilite (üreme yeteneği)”, which is a gender-neutral counterpart for fertility.

In Segment 5 of the MT output, the phrase “hayvanlarda yapılan bulgulara göre” is a problematic one because it refers to “based on findings made in animals”. Findings are not made, they are obtained. Therefore, it could have been replaced with alternatives such as: “hayvanlardan edinilen bulgulara göre”, or “hayvanlarda yapılan çalışmaların bulgularına göre”, or even “hayvanlardaki bulgulara göre” (and so on).

Unlike the reference text, Google’s MT engine translates “male infertility” as “erkek kısırlığı” rather than “erkek infertilitesi” which is a way of simplification, as can be seen in Segment 5. In parantheses, the collocation is explained as “(düşük veya hiç sperm sayısı)” which leads to an error. An alternative would be “düşük sperm sayısı veya hiç sperm sayısı olmaması”.

Segment 9 has a grammatical error which stems from problematic use of parts of speech: the phrase “You may feel tired or dizzy after taking Venclyxto” has been translated as “Venclyxto aldıktan sonra yorgun veya baş dönmesi hissedebilirsiniz”. The problem here lies in the collocation “yorgun veya baş dönmesi hissedebilirsiniz” which actually means “you may feel tired or dizziness”. “Yorgun” in Turkish is an adjective whereas “baş dönmesi” is a noun phrase. Both parts of speech should have either been adjectives or nouns. To overcome this issue, a correct alternative would be “yorgunluk veya baş dönmesi” hissedebilirsiniz by turning the term “yorgun” into a noun (tired to tiredness).

5.3.3.2 GPT-4 – zero-shot

Table 30: ST-2, Reference Text for ST-2, and MT Output (GPT-4 – zero-shot)

Segment No.	Source Text:	Reference Text:	MT Output:
1.	Breast-feeding	Emzirme	Emzirme
2.	Do not breast-feed while you are taking this medicine.	Bu ilacı kullanırken emzirmeyiniz.	Bu ilacı alırken emzirmeyin.
3.	It is not known whether the active substance in Venclyxto passes into breast milk.	VENCLYXTO’nun etkin maddesinin anne sütüne geçip geçmediği bilinmemektedir.	Venclyxto’nun etkin maddesinin anne sütüne geçip geçmediği bilinmemektedir.
4.	Fertility	Fertilite (üreme yeteneği)	Üreme Yeteneği
5.	Based on findings in animals, Venclyxto may cause male infertility (low or no sperm count).	Hayvan çalışmaları bulgularına göre, VENCLYXTO erkek infertilitesine (sperm sayısında düşme ya da sperm yokluğu) neden olabilir.	Hayvanlar üzerinde yapılan bulgulara dayanarak, Venclyxto erkeklerde kısırlığa (düşük veya hiç sperm sayısı) neden olabilir.
6.	This may affect your ability to father a child.	Bu durum baba olma yetinizi etkileyebilir.	Bu, çocuk sahibi olma yeteneğinizi etkileyebilir.
7.	Ask your doctor for advice on sperm storage before starting treatment with Venclyxto.	VENCLYXTO tedavisine başlamadan önce sperm depolama hakkında doktorunuza danışınız.	Venclyxto ile tedaviye başlamadan önce doktorunuza spermi saklama konusunda danışın.
8.	Driving and using machines	Araç ve makine kullanımı	Araç ve Makine Kullanımı

Table 30 - continue

9.	You may feel tired or dizzy after taking Venclyxto, which may affect your ability to drive or use tools or machines.	VENCLYXTO kullandıktan sonra, araç ve makine kullanımı	Venclyxto'yu aldıktan sonra yorgunluk veya baş dönmesi hissedebilirsiniz, bu da araç kullanma veya alet ve makine
		kabiliyetinizi etkileyebilecek yorgunluk veya baş dönmesi hissedebilirsiniz.	kullanma yeteneğinizi etkileyebilir.

- Error-typology chart of the MT output:

Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use				0	0
Terminology	Incorrect term use	1	2		3	7
Accuracy	Mistranslation				0	0
Accuracy	Overtranslation				0	0
Accuracy	Undertranslation	1	1		2	4
Accuracy	Addition				0	0
Accuracy	Omission				0	0
Accuracy	Untranslated				0	0
Linguistic Conventions	Grammar	3			3	3
Linguistic Conventions	Spelling				0	0
Syntax	Syntactical errors				0	0
Style & Clarity	Unnatural Phrasing	1	1		2	4
Style & Clarity	Poor Readability				0	0
Custom	Layout, character & other				0	0
		Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
		6	4	0	10	18

Figure 18: Error-typology Chart (GPT-4 – zero-shot)

In this output, when we look at Segment 4, we see that the sub-heading “Fertility” is oversimplified as “Üreme Yeteneği”, which is a purified substitute for fertility. As can be seen in the reference text, the correct way to put it would have been “Fertilite (Üreme Yeteneği)” by purifying the term in parantheses. Leaving the sub-heading solely as “üreme yeteneği” leads to an ambiguity.

GPT-4’s raw output seems to have rendered “ability to father a child” in Segment 6 as “çocuk sahibi olma yeteneğinizi” which is another error of incorrect term use. This is mainly because ability is an inherent occurrence whereas “yetenek” in Turkish corresponds to the concept of “skill” or “talent”, which are customized ways of performing inherent potential. That is to say, “yeti” and “yetenek” are different concepts based on the nature of humankind. “Yetenek” (skill/talent) is how humanbeings use their inherent “yeti” (ability) as a tool. Therefore, this term is not suitable for such usage given that fathering a child is an inherent biological occurrence.

Segment 7 has an unnatural phrasing error stemming from the translation of “sperm storage” which is translated as “spermi saklama”. In Turkish language, the act of “saklama” often signifies the act of “preserving” or “hiding” although ST intends to refer the reader to a physician on asking for consultation on how to “store” sperm due to potential impact of a drug that can cause low or no sperm count in males. “Depolama” is the correct counterpart for the act of “storing”.

5.3.3.3 GPT-4 – RBP

Table 31: ST-2, Reference Text for ST-2, and MT Output (GPT-4 - RBP)

Segment No.	Source Text:	Reference Text:	MT Output:
1.	Breast-feeding	Emzirme	Emzirme
2.	Do not breast-feed while you are taking this medicine.	Bu ilacı kullanırken emzirmeyiniz.	Bu ilacı kullanırken emzirmeyiniz.
3.	It is not known whether the active substance in Venclyxto passes into breast milk.	VENCLYXTO’nun etkin maddesinin anne sütüne geçip geçmediği bilinmemektedir.	Venclyxto’nun etkin maddesinin anne sütüne geçip geçmediği bilinmemektedir.
4.	Fertility	Fertilite (üreme yeteneği)	Fertilite

Table 31 - continue

5.	Based on findings in animals, Venclyxto may cause male infertility (low or no sperm count).	Hayvan alıřmaları bulgularına gre, VENCLYXTO erkek infertilitesine (sperm sayısında dřme ya da sperm yokluęu) neden olabilir.	Hayvanlar zerinde yapılan bulgulara dayanarak, Venclyxto erkeklerde infertiliteye (dřk veya hi sperm sayısı) neden olabilir.
6.	This may affect your ability to father a child.	Bu durum baba olma yetinizi etkileyebilir.	Bu, ocuk sahibi olma yeteneęinizi etkileyebilir.
7.	Ask your doctor for advice on sperm storage before starting treatment with Venclyxto.	VENCLYXTO tedavisine bařlamadan nce sperm depolama hakkında doktorunuza danıřınız.	Venclyxto ile tedaviye bařlamadan nce doktorunuza sperminizi saklama konusunda danıřın.
8.	Driving and using machines	Ara ve makine kullanımı	Ara ve makine kullanımı
9.	You may feel tired or dizzy after taking Venclyxto, which may affect your ability to drive or use tools or machines.	VENCLYXTO kullandıktan sonra, ara ve makine kullanımı kabiliyetinizi etkileyebilecek yorgunluk veya bař dnmesi hissedebilirsiniz.	Venclyxto aldıktan sonra yorgunluk veya bař dnmesi hissedebilirsiniz, bu da ara kullanma veya alet ve makine kullanma yeteneęinizi etkileyebilir.

- Error-typology chart of the MT output:

Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use				0	0
Terminology	Incorrect term use	1	2		3	7
Accuracy	Mistranslation				0	0
Accuracy	Overtranslation				0	0
Accuracy	Undertranslation				0	0
Accuracy	Addition				0	0
Accuracy	Omission				0	0
Accuracy	Untranslated				0	0
Linguistic Conventions	Grammar	1			1	1
Linguistic Conventions	Spelling				0	0
Syntax	Syntactical errors				0	0
Style & Clarity	Unnatural Phrasing	1	1		2	4
Style & Clarity	Poor Readability				0	0
Custom	Layout, character & other				0	0
		Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
		3	3	0	6	12

Figure 19: Error-typology Chart (GPT-4 - RBP)

5.3.3.4 Gemini-2.0-flash – zero-shot

Table 32: ST-2, Reference Text for ST-2, and MT Output (Gemini-2.0-flash – zero-shot)

Segment No.	Source Text:	Reference Text:	MT Output:
1.	Breast-feeding	Emzirme	Emzirme
2.	Do not breast-feed while you are taking this medicine.	Bu ilacı kullanırken emzirmeyiniz.	Bu ilacı alırken emzirmeyin.
3.	It is not known whether the active substance in Venclyxto passes into breast milk.	VENCLYXTO'nun etkin maddesinin anne sütüne geçip geçmediği bilinmemektedir.	Venclyxto'daki aktif maddenin anne sütüne geçip geçmediği bilinmemektedir.
4.	Fertility	Fertilite (üreme yeteneği)	Doğurganlık

Table 32 - continue

5.	Based on findings in animals, Venclyxto may cause male infertility (low or no sperm count).	Hayvan çalışmaları bulgularına göre, VENCLYXTO erkek infertilitesine (sperm sayısında düşme ya da sperm yokluğu) neden olabilir.	Hayvanlarda yapılan bulgulara dayanarak, Venclyxto erkeklerde kısırlığa (düşük veya hiç sperm sayısı) neden olabilir.
6.	This may affect your ability to father a child.	Bu durum baba olma yetinizi etkileyebilir.	Bu, çocuk sahibi olma yeteneğinizi etkileyebilir.
7.	Ask your doctor for advice on sperm storage before starting treatment with Venclyxto.	VENCLYXTO tedavisine başlamadan önce sperm depolama hakkında doktorunuza danışınız.	Venclyxto ile tedaviye başlamadan önce doktorunuzdan sperm saklama konusunda tavsiye isteyin.
8.	Driving and using machines	Araç ve makine kullanımı	Araç ve makine kullanımı
9.	You may feel tired or dizzy after taking Venclyxto, which may	VENCLYXTO kullandıktan sonra, araç ve makine kullanımı	Venclyxto aldıktan sonra yorgun veya başınız dönmüş hissedebilirsiniz, bu da araç.
	affect your ability to drive or use tools or machines.	kabiliyetinizi etkileyebilecek yorgunluk veya baş dönmesi hissedebilirsiniz.	veya alet veya makine kullanma yeteneğinizi etkileyebilir

- Error-typology chart of the MT output:

Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use				0	0
Terminology	Incorrect term use	1	3		4	10
Accuracy	Mistranslation				0	0
Accuracy	Overtranslation				0	0
Accuracy	Undertranslation	1		1	2	6
Accuracy	Addition				0	0
Accuracy	Omission				0	0
Accuracy	Untranslated				0	0
Linguistic Conventions	Grammar	2			2	2
Linguistic Conventions	Spelling				0	0
Syntax	Syntactical errors				0	0
Style & Clarity	Unnatural Phrasing	2	1		3	5
Style & Clarity	Poor Readability				0	0
Custom	Layout, character & other				0	0
		Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
		6	4	1	11	23

Figure 20: Error-typology Chart (Gemini-2.0-flash – zero-shot)

With a total of 23 points, this output generated by Gemini’s 2.0-flash model (with raw prompt) possesses the highest TQE penalty score, which makes it the most problematic output, from a qualitative approach.

Besides the recurring errors across all outputs, there are two distinct errors unique to this output: first one worth mentioning is that, on Segment 3, the term “active substance” was rendered as “aktif madde” whereas in the literature of pharmaceuticals, there is an established use of “etkin madde”.

The second one lies in the instructions on Segment 9: the phrase “Venclyxto aldıktan sonra yorgun veya başınız dönmüş hissedebilirsiniz...” has a problematic denotation in the sub-clause “başınız dönmüş hissedebilirsiniz”, which refers to feeling “as if” dizzy. This implies a subjective feeling of dizziness rather than a direct physiological effect, whereas the instructions clearly emphasize the possible dizziness a patient can experience after taking the medication in question.

5.3.3.5 Gemini-2.0-flash – RBP

Table 33: ST-2, Reference Text for ST-2, and MT Output (Gemini-2.0-flash - RBP)

Segment No.	Source Text:	Reference Text:	MT Output:

Table 33 - continue

1.	Breast-feeding	Emzirme	Emzirme
2.	Do not breast-feed while you are taking this medicine.	Bu ilacı kullanırken emzirmeyiniz.	Bu ilacı alırken emzirmeyiniz.
3.	It is not known whether the active substance in Venclyxto passes into breast milk.	VENCLYXTO'nun etkin maddesinin anne sütüne geçip geçmediği bilinmemektedir.	Venclyxto'daki etkin maddenin anne sütüne geçip geçmediği bilinmemektedir.
4.	Fertility	Fertilite (üreme yeteneği)	Doğurganlık
5.	Based on findings in animals, Venclyxto may cause male infertility (low or no sperm count).	Hayvan çalışmaları bulgularına göre, VENCLYXTO erkek infertilitesine (sperm sayısında düşme ya da sperm yokluğu) neden olabilir.	Hayvanlardaki bulgulara dayanarak, Venclyxto erkeklerde kısırlığa (düşük veya hiç sperm sayısı) neden olabilir.
6.	This may affect your ability to father a child.	Bu durum baba olma yetinizi etkileyebilir.	Bu, çocuk sahibi olma yeteneğinizi etkileyebilir.
7.	Ask your doctor for advice on sperm storage before starting treatment with Venclyxto.	VENCLYXTO tedavisine başlamadan önce sperm depolama hakkında doktorunuza danışınız.	Venclyxto ile tedaviye başlamadan önce doktorunuzdan sperm saklama konusunda tavsiye isteyiniz.
8.	Driving and using machines	Araç ve makine kullanımı	Araç ve Makine Kullanımı

Table 33 - continue

9.	You may feel tired or dizzy after taking Venclyxto, which may affect your ability to drive or use tools or machines.	VENCLYXTO kullandıktan sonra, araç ve makine kullanımı kabiliyetinizi etkileyebilecek yorgunluk veya baş dönmesi hissedebilirsiniz.	Venclyxto aldıktan sonra yorgun veya başınız dönmüş hissedebilirsiniz, bu da araç veya alet veya makine kullanma yeteneğinizi etkileyebilir.
----	--	--	--

- Error-typology chart of the MT output:

Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use				0	0
Terminology	Incorrect term use	1	2		3	7
Accuracy	Mistranslation				0	0
Accuracy	Overtranslation				0	0
Accuracy	Undertranslation	1		1	2	6
Accuracy	Addition				0	0
Accuracy	Omission				0	0
Accuracy	Untranslated				0	0
Linguistic Conventions	Grammar				0	0
Linguistic Conventions	Spelling				0	0
Syntax	Syntactical errors				0	0
Style & Clarity	Unnatural Phrasing	1	1		2	4
Style & Clarity	Poor Readability				0	0
Custom	Layout, character & other				0	0
		Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
		3	3	1	7	17

Figure 21: Error-typology Chart (Gemini-2.0-flash - RBP)

5.3.3.6 DeepSeek – R1 – zero-shot

Table 34: ST-2, Reference Text for ST-2, and MT Output (DeepSeek-R1 – zero-shot)

Segment No.	Source Text:	Reference Text:	MT Output:
1.	Breast-feeding	Emzirme	Emzirme

Table 34 - continue

2.	Do not breast-feed while you are taking this medicine.	Bu ilacı kullanırken emzirmeyiniz.	Bu ilacı kullanırken emzirmeyin.
3.	It is not known whether the active substance in Venclyxto passes into breast milk.	VENCLYXTO'nun etkin maddesinin anne sütüne geçip geçmediği bilinmemektedir.	Venclyxto'nun etkin maddesinin anne sütüne geçip geçmediği bilinmemektedir.
4.	Fertility	Fertilite (üreme yeteneği)	Üreme Yeteneği
5.	Based on findings in animals, Venclyxto may cause male infertility (low or no sperm count).	Hayvan çalışmaları bulgularına göre, VENCLYXTO erkek infertilitesine (sperm sayısında düşme ya da sperm yokluğu) neden olabilir.	Hayvanlar üzerinde yapılan çalışmalara göre, Venclyxto erkeklerde kısırlığa (düşük sperm sayısı veya hiç sperm olmaması) neden olabilir.
6.	This may affect your ability to father a child.	Bu durum baba olma yetinizi etkileyebilir.	Bu durum, çocuk sahibi olma yeteneğinizi etkileyebilir.
7.	Ask your doctor for advice on sperm storage before starting treatment with Venclyxto.	VENCLYXTO tedavisine başlamadan önce sperm depolama hakkında doktorunuza danışınız.	Venclyxto tedavisine başlamadan önce sperm depolama konusunda doktorunuza danışın.
8.	Driving and using machines	Araç ve makine kullanımı	Araç ve Makine Kullanımı
9.	You may feel tired or dizzy after taking Venclyxto, which may affect your ability to drive or use tools or machines.	VENCLYXTO kullandıktan sonra, araç ve makine kullanımı kabiliyetinizi etkileyebilecek yorgunluk veya baş dönmesi hissedebilirsiniz.	Venclyxto aldıktan sonra kendinizi yorgun veya başınız dönmüş hissedebilirsiniz. Bu durum, araç kullanma veya alet/makine kullanma yeteneğinizi etkileyebilir

- Error-typology chart of the MT output:

Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use				0	0
Terminology	Incorrect term use	1	1		2	4
Accuracy	Mistranslation				0	0
Accuracy	Overtranslation				0	0
Accuracy	Undertranslation	1	1		2	4
Accuracy	Addition				0	0
Accuracy	Omission				0	0
Accuracy	Untranslated				0	0
Linguistic Conventions	Grammar	2			2	2
Linguistic Conventions	Spelling				0	0
Syntax	Syntactical errors				0	0
Style & Clarity	Unnatural Phrasing				0	0
Style & Clarity	Poor Readability				0	0
Custom	Layout, character & other				0	0
		Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
		4	2	0	6	10

Figure 22: Error-typology Chart (DeepSeek-R1 – zero-shot)

5.3.3.7 DeepSeek – R1 – RBP

Table 35: ST-2, Reference Text for ST-2, and MT Output (DeepSeek-R1 – RBP)

Segment No.	Source Text:	Reference Text:	MT Output:
1.	Breast-feeding	Emzirme	Emzirme
2.	Do not breast-feed while you are taking this medicine.	Bu ilacı kullanırken emzirmeyiniz.	Venclyxto’yu kullanırken emzirmeyin.
3.	It is not known whether the active substance in Venclyxto passes into breast milk.	VENCLYXTO’nun etkin maddesinin anne sütüne geçip geçmediği bilinmemektedir.	Venclyxto’nun etkin maddesinin anne sütüne geçip geçmediği bilinmemektedir.
4.	Fertility	Fertilite (üreme yeteneği)	Doğurganlık
5.	Based on findings in animals, Venclyxto may cause male infertility (low or no sperm count).	Hayvan çalışmaları bulgularına göre, VENCLYXTO erkek infertilitesine (sperm sayısında düşme ya da	Hayvanlar üzerinde yapılan çalışmalara göre, Venclyxto erkeklerde kısırlığa (düşük sperm sayısı veya hiç sperm olmaması) neden olabilir.

Table 35 - continue

		sperm yokluğu) neden olabilir.	
6.	This may affect your ability to father a child.	Bu durum baba olma yetinizi etkileyebilir.	Bu durum, çocuk sahibi olma yeteneğinizi etkileyebilir.
7.	Ask your doctor for advice on sperm storage before starting treatment with Venclyxto.	VENCLYXTO tedavisine başlamadan önce sperm depolama hakkında doktorunuza danışınız.	Venclyxto tedavisine başlamadan önce sperm depolama konusunda doktorunuza danışın.
8.	Driving and using machines	Araç ve makine kullanımı	Araç ve Makine Kullanımı
9.	You may feel tired or dizzy after taking Venclyxto, which may affect your ability to drive or use tools or machines.	VENCLYXTO kullandıktan sonra, araç ve makine kullanımı kabiliyetinizi etkileyebilecek yorgunluk veya baş dönmesi hissedebilirsiniz.	Venclyxto aldıktan sonra kendinizi yorgun hissedebilir veya baş dönmesi yaşayabilirsiniz. Bu durum, araç kullanma ya da alet/makine kullanma becerinizi etkileyebilir.

o Error-typology chart of the MT output:

Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use				0	0
Terminology	Incorrect term use		1		1	3
Accuracy	Mistranslation				0	0
Accuracy	Overtranslation				0	0
Accuracy	Undertranslation	1		1	2	6
Accuracy	Addition				0	0
Accuracy	Omission				0	0
Accuracy	Untranslated				0	0
Linguistic Conventions	Grammar	2			2	2
Linguistic Conventions	Spelling				0	0
Syntax	Syntactical errors				0	0
Style & Clarity	Unnatural Phrasing				0	0
Style & Clarity	Poor Readability				0	0
Custom	Layout, character & other				0	0
		Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
		3	1	1	5	11

Figure 23: Error-typology Chart (DeepSeek-R1 - RBP)

5.3.3.8 Claude-3.5-Sonnet – zero-shot

Table 36: ST-2, Reference Text for ST-2, and MT Output (Claude-3.5-Sonnet – zero-shot)

Segment No.	Source Text:	Reference Text:	MT Output:
1.	Breast-feeding	Emzirme	Emzirme
2.	Do not breast-feed while you are taking this medicine.	Bu ilacı kullanırken emzirmeyiniz.	Bu ilacı kullanırken emzirmeyin.
3.	It is not known whether the active substance in Venclyxto passes into breast milk.	VENCLYXTO'nun etkin maddesinin anne sütüne geçip geçmediği bilinmemektedir.	Venclyxto'daki etkin maddenin anne sütüne geçip geçmediği bilinmemektedir.
4.	Fertility	Fertilite (üreme yeteneği)	Doğurganlık
5.	Based on findings in animals, Venclyxto may cause male infertility (low or no sperm count).	Hayvan çalışmaları bulgularına göre, VENCLYXTO erkek infertilitesine (sperm sayısında düşme ya da sperm yokluğu) neden olabilir.	Hayvanlarda yapılan çalışmalara dayanarak, Venclyxto erkek kısırlığına (düşük veya sıfır sperm sayısı) neden olabilir.
6.	This may affect your ability to father a child.	Bu durum baba olma yetinizi etkileyebilir.	Bu, çocuk sahibi olma yeteneğinizi etkileyebilir.
7.	Ask your doctor for advice on sperm storage before starting treatment with Venclyxto.	VENCLYXTO tedavisine başlamadan önce sperm depolama hakkında doktorunuza danışınız.	Venclyxto ile tedaviye başlamadan önce sperm depolama konusunda doktorunuza danışın.
8.	Driving and using machines	Araç ve makine kullanımı	Araç ve makine kullanımı
9.	You may feel tired or dizzy after taking Venclyxto, which may	VENCLYXTO kullandıktan sonra, araç ve	Venclyxto aldıktan sonra yorgunluk veya baş dönmesi

Table 36 - continue

	affect your ability to drive or use tools or machines.	makine kullanımı kabiliyetinizi etkileyebilecek yorgunluk veya baş dönmesi hissedebilirsiniz.	hissedebilirsiniz; bu durum araç veya alet ya da makine kullanma yeteneğinizi etkileyebilir.
--	--	---	--

- Error-typology chart of the MT output:

Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use				0	0
Terminology	Incorrect term use	1	1		2	4
Accuracy	Mistranslation				0	0
Accuracy	Overtranslation				0	0
Accuracy	Undertranslation	1		1	2	6
Accuracy	Addition				0	0
Accuracy	Omission				0	0
Accuracy	Untranslated				0	0
Linguistic Conventions	Grammar	2			2	2
Linguistic Conventions	Spelling				0	0
Syntax	Syntactical errors				0	0
Style & Clarity	Unnatural Phrasing				0	0
Style & Clarity	Poor Readability				0	0
Custom	Layout, character & other				0	0
		Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
		4	1	1	6	12

Figure 24: Error-typology Chart (Claude-3.5-Sonnet – zero-shot)

5.3.3.9 Claude-3.5-Sonnet – RBP

Table 37: ST-2, Reference Text for ST-2, and MT Output (Claude-3.5-Sonnet - RBP)

Segment No.	Source Text:	Reference Text:	MT Output:
1.	Breast-feeding	Emzirme	Emzirme
2.	Do not breast-feed while you are taking this medicine.	Bu ilacı kullanırken emzirmeyiniz.	Bu ilacı aldığınız süre boyunca emzirmeyiniz.

Table 37 - continue

3.	It is not known whether the active substance in Venclyxto passes into breast milk.	VENCLYXTO'nun etkin maddesinin anne sütüne geçip geçmediği bilinmemektedir.	Venclyxto'nun etkin maddesinin anne sütüne geçip geçmediği bilinmemektedir.
4.	Fertility	Fertilite (üreme yeteneği)	Doğurganlık
5.	Based on findings in animals, Venclyxto may cause male infertility (low or no sperm count).	Hayvan çalışmaları bulgularına göre, VENCLYXTO erkek infertilitesine (sperm sayısında düşme ya da sperm yokluğu) neden olabilir.	Hayvanlar üzerinde yapılan çalışmalara dayanarak, Venclyxto erkek doğurganlığında azalmaya (düşük sperm sayısı veya sperm yokluğu) neden olabilir.
6.	This may affect your ability to father a child.	Bu durum baba olma yetinizi etkileyebilir.	Bu durum çocuk sahibi olma yeteneğinizi etkileyebilir.
7.	Ask your doctor for advice on sperm storage before starting treatment with Venclyxto.	VENCLYXTO tedavisine başlamadan önce sperm depolama hakkında doktorunuza danışınız.	Venclyxto ile tedaviye başlamadan önce sperm saklama konusunda doktorunuzdan tavsiye alınız.
8.	Driving and using machines	Araç ve makine kullanımı	Araç ve makine kullanımı
9.	You may feel tired or dizzy after taking Venclyxto, which may affect your ability to drive or use tools or machines.	VENCLYXTO kullandıktan sonra, araç ve makine kullanımı kabiliyetinizi etkileyebilecek yorgunluk veya baş dönmesi hissedebilirsiniz.	Venclyxto aldıktan sonra yorgunluk veya baş dönmesi hissedebilirsiniz; bu durum araç veya alet ya da makine kullanma yeteneğinizi etkileyebilir.

- Error-typology chart of the MT output:

Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use				0	0
Terminology	Incorrect term use	1	2		3	7
Accuracy	Mistranslation			1	1	5
Accuracy	Overtranslation				0	0
Accuracy	Undertranslation			1	1	5
Accuracy	Addition				0	0
Accuracy	Omission				0	0
Accuracy	Untranslated				0	0
Linguistic Conventions	Grammar				0	0
Linguistic Conventions	Spelling				0	0
Syntax	Syntactical errors				0	0
Style & Clarity	Unnatural Phrasing	1			1	1
Style & Clarity	Poor Readability				0	0
Custom	Layout, character & other				0	0
		Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
		2	2	2	6	18

Figure 25: Error-typology Chart (Claude-3.5-Sonnet - RBP)

Role-based prompted model of Claude exhibits a critical error on Segment 5: “...Venclyxto erkek doğurganlığında azalmaya (..) neden olabilir.” which signifies “decrease in childbearing potential” rather than male fertility. The concept of “fertility” in English language is a gender-neutral concept, whereas the concept of “doğurganlık” in Turkish language signifies childbearing potential and is a phenomenon pertaining to biologically female individuals, hence the usage in Segment 5 is an erroneous one as the collocation in the output refers to childbearing in males, which is a critical mistranslation leading to ambiguity.

5.3.4 Overall Results Overview for ST-2’s TT Outputs

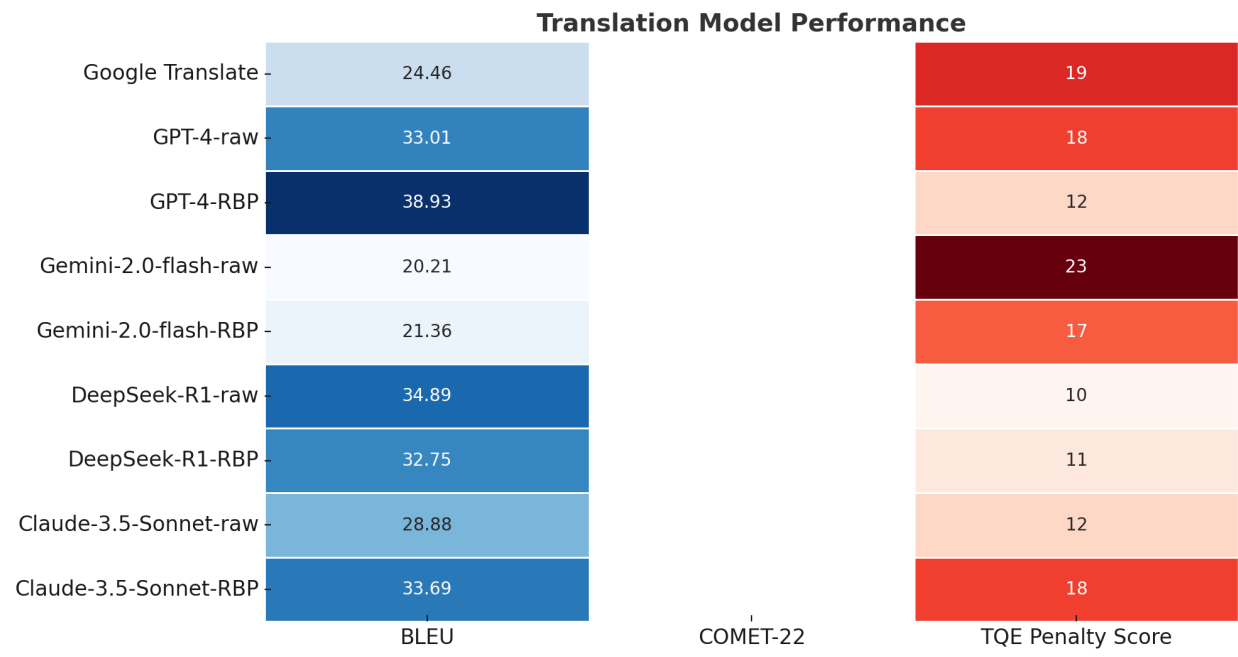


Figure 26: Heatmap Overview of Quantitative and Qualitative Metric Results of ST-2’s TT Outputs

5.3.5 Analysis on ST-3’s TT Outputs

This subsection will review 9 MT outputs against ST-3. As in the previous analysis on ST-2’s TT outputs, I will not elaborate on every individual error however will conduct a commentary on fundamental errors.

5.3.5.1 Google Translate – raw

Table 38: ST-3, Reference Text for ST-3, and MT Output (Google Translate - raw)

Segment No.	Source Text:	Reference Text:	MT Output:
1.	STEP 2	ADIM 2	ADIM 2

Table 38 - continue

2.	Place the following items on a clean, flat surface:	Aşağıdaki malzemeleri temiz ve düz bir yüzey üzerine yerleştiriniz:	Aşağıdaki maddeleri temiz ve düz bir yüzeye koyun:
3.	• 1 pre-filled pen	• 1 adet kullanıma hazır kalem	• 1 önceden doldurulmuş kalem
4.	• 1 alcohol pad (not included in the carton)	• 1 adet alkollü ped (kutuya dahil değildir)	• 1 alkol pedi (kartona dahil değildir)
5.	• 1 cotton ball or gauze pad (not included in the carton)	• 1 adet pamuk yumağı veya gazlı bez (kutuya dahil değildir)	• 1 pamuk topu veya gazlı bez pedi (kartona dahil değildir)
6.	• special disposal container (not included in the carton)	• Özel atık kabı (kutuya dahil değildir)	• özel atık kabı (kartona dahil değildir)
7.	Wash and dry your hands.	Ellerinizi yıkayıp kurulayınız.	Ellerinizi yıkayın ve kurulayın.
8.	STEP 3	ADIM 3	ADIM 3
9.	Choose from these 3 areas to inject:	Enjeksiyon için şu 3 bölgeden birini seçiniz:	Enjeksiyon yapmak için şu 3 bölgeden birini seçin:
10.	• front of left thigh	• sol uyluğun önü	• sol uyluğun önü
11.	• front of right thigh	• sağ uyluğun önü	• sağ uyluğun önü
12.	• your belly (abdomen) at least 5 cm from your belly button (navel)	• göbek deliğinin en az 5 cm uzağında olmak üzere göbek bölgesi (karın)	• göbeğinizden (karın) en az 5 cm uzakta
13.	Before the injection, wipe where you will inject in a circular motion with an alcohol pad.	Enjeksiyondan önce, enjeksiyon yapılacak yeri bir alkollü ped ile dairesel hareketlerle siliniz.	Enjeksiyondan önce, enjeksiyon yapacağınız yeri alkollü bir pedle dairesel hareketlerle silin.
14.	• Do not touch or blow on the injection site after it is cleaned.	• Temizledikten sonra enjeksiyon bölgesine dokunmayınız ya da üfleme yapınız.	Temizlendikten sonra enjeksiyon bölgesine dokunmayın veya üfleme yapmayın.

Table 38 - continue

15.	Allow the skin to dry before injecting	Enjeksiyondan önce derinizin kurumasını bekleyiniz.	Enjekte etmeden önce cildin kurumasını bekleyin
16.	• Do not inject through clothes	• Kıyafetlerin üstünden enjeksiyon yapmayınız.	• Giysilerinizin üzerinden enjeksiyon yapmayın
17.	• Do not inject into skin that is sore, bruised, red, hard, scarred, or has stretch marks	• Ağrıyan, morarmış, kızarmış, sertleşmiş, yaralı veya çatlak izleri olan deri üzerine enjeksiyon yapmayınız.	• Ağrılı, morarmış, kırmızı, sert, yara izi olan veya çatlak izleri olan cilde enjeksiyon yapmayın
18.	• Do not inject into areas affected by psoriasis	• Sedef hastalığından etkilenmiş bölgelere enjeksiyon yapmayınız.	• Sedef hastalığından etkilenen bölgelere enjeksiyon yapmayın

○ Error-typology chart of the MT output:

Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use	1			1	1
Terminology	Incorrect term use	1	4		5	13
Accuracy	Mistranslation				0	0
Accuracy	Overtranslation				0	0
Accuracy	Undertranslation				0	0
Accuracy	Addition	1			1	1
Accuracy	Omission	3		1	4	8
Accuracy	Untranslated				0	0
Linguistic Conve	Grammar				0	0
Linguistic Conve	Spelling				0	0
Syntax	Syntactical errors				0	0
Style & Clarity	Unnatural Phrasing	1	1		2	4
Style & Clarity	Poor Readability			1	1	5
Custom	Layout, character & other				0	0
		Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
		7	5	2	14	32

Figure 27: Error-typology Chart (Google Translate - raw)

Among the 9 outputs, this output is the 7th ranked output in terms of BLEU, and the 9th output in terms of TQE penalty score evaluation. Besides minor errors, there are 5 major, 2 critical errors noted, which can cause serious ambiguities, hence, misleading instructions that jeopardize effective use of a pharmaceutical product. Such findings are crucial when patient safety and health is at stake.

To delve into the analysis, in Segment 2, we can see that the sub-clause “place the following items” was translated as “aşağıdaki maddeleri (...) koyun”. The term “item” does not sound natural when it is translated as “madde”, moreover, the items listed could not be deemed “madde” in this context. Although a minor error, this is an incorrect term use.

In Segments 4, 5, 6: the term “carton” in parentheses is an item that contains pharmaceutical products alongside their package leaflets. It is essentially a package. However, the term “karton”, which is the equivalent in the MT output, is not an accurate counterpart. “Karton” is a material that a package can be made of, whereas the output rendered it as if it was the entire package. The correct counterpart for carton would be “kutu”, as can be seen in the reference Segments 4, 5, 6.

One of the most crucial parts of the source extract is the instructions offered in Segments 10, 11, 12 as they describe injection sites for the pre-filled pen, which is the pharmaceutical product in question. When one looks at Segment 12, it can be inferred that the abdomen area with at least 5 cm distance to belly button is one of the appropriate injection sites for the pre-filled pen. Then, in the Turkish output, it is rendered as “göbeğinizden (karın) en az 5 cm uzakta” which has been marked as critical error of poor readability because the phrase fails to clearly describe the specific point of reference -i.e., 5 cm from where- thereby hindering the clarity of the instruction. This is a remarkably vital misguidance for the reader as it concerns one of the areas that a patient will perform injection on. Plus, the term “belly button” is the counterpart for “göbek deliği” in Turkish, yet, there is no such term seen in the output, which is an omission error that is critical by severity of the error. In order for the reader to perceive the injection site precisely, this is a vital aspect where omission must not be made. In Segment 4, “alcohol pad” was rendered by Google Translate’s MT output as

“alkol pedi” whereas in Segment 13 it is rendered as “alkollü bir pedle”, which is a case of inconsistency.

As for the last observation to put forth, Segment 17 describes cases where the injection shouldn’t be performed: “Do not inject into skin that is sore, bruised, red, hard, scarred, or has stretch marks”. A skin being red does not necessarily bear the same denotation in Turkish language as it does it in English. Its denotative equivalent is “kırmızı” in Turkish however when it comes to skin, it would be either “kızarıklık” or “kızarmış”, although the MT output directly puts it as “kırmızı”. As this can also lead to ambiguities, this is an incorrect term use error that is major by severity. I haven’t elaborated on every unique error in this subsection, although this is the most erroneous output together with the errors excluded from the scope of the commentary.

5.3.5.2 GPT-4 – zero-shot

Table 39: ST-3, Reference Text for ST-3, and MT Output (GPT-4 – zero-shot)

Segment No.	Source Text:	Reference Text:	MT Output:
1.	STEP 2	ADIM 2	ADIM 2
2.	Place the following items on a clean, flat surface:	Aşağıdaki malzemeleri temiz ve düz bir yüzey üzerine yerleştiriniz:	Aşağıdaki eşyaları temiz, düz bir yüzeye yerleştirin:
3.	• 1 pre-filled pen	• 1 adet kullanıma hazır kalem	• 1 adet dolu kalem
4.	• 1 alcohol pad (not included in the carton)	• 1 adet alkollü ped (kutuya dahil değildir)	• 1 adet alkol pedi (kutuya dahil değildir)
5.	• 1 cotton ball or gauze pad (not included in the carton)	• 1 adet pamuk yumağı veya gazlı bez (kutuya dahil değildir)	• 1 adet pamuk topu veya gazlı bez (kutuya dahil değildir)
6.	• special disposal container (not included in the carton)	• Özel atık kabı (kutuya dahil değildir)	• özel atık konteyneri (kutuya dahil değildir)

Table 39 - continue

7.	Wash and dry your hands.	Ellerinizi yıkayıp kurulayınız.	Ellerinizi yıkayın ve kurulayın.
8.	STEP 3	ADIM 3	ADIM 3
9.	Choose from these 3 areas to inject:	Enjeksiyon için şu 3 bölgeden birini seçiniz:	Enjeksiyon yapmak için bu 3 bölgeden birini seçin:
10.	• front of left thigh	• sol uyluğun önü	• sol uyluğun önü
11.	• front of right thigh	• sağ uyluğun önü	• sağ uyluğun önü
12.	• your belly (abdomen) at least 5 cm from your belly button (navel)	• göbek deliğinin en az 5 cm uzağında olmak üzere göbek bölgesi (karın)	• göbeğiniz (karın), göbek düğmenizden (navel) en az 5 cm uzaklıkta
13.	Before the injection, wipe where you will inject in a circular motion with an alcohol pad.	Enjeksiyondan önce, enjeksiyon yapılacak yeri bir alkollü ped ile dairesel hareketlerle siliniz.	Enjeksiyon yapmadan önce, enjeksiyon yapılacak yeri alkol pedi ile dairesel hareketlerle silin.
14.	• Do not touch or blow on the injection site after it is cleaned.	• Temizledikten sonra enjeksiyon bölgesine dokunmayınız ya da üfleme yapınız.	• Temizlendikten sonra enjeksiyon yerine dokunmayın veya üfleme yapmayın.
15.	Allow the skin to dry before injecting	Enjeksiyondan önce derinizin kurumasını bekleyiniz.	Enjekte etmeden önce cildin kurumasını bekleyin
16.	• Do not inject through clothes	• Kıyafetlerin üstünden enjeksiyon yapmayınız .	• Kıyafetlerin üzerinden enjekte etmeyin
17.	• Do not inject into skin that is sore, bruised, red, hard, scarred, or has stretch marks	• Ağrıyan, morarmış, kızarmış, sertleşmiş, yaralı veya çatlak izleri olan deri üzerine enjeksiyon yapmayınız .	• Ağrılı, morarmış, kırmızı, sert, yara izli veya çatlaklı ciltlere enjekte etmeyin

Table 39 - continue

18.	• Do not inject into areas affected by psoriasis	• Sedef hastalığından etkilenmiş bölgelere enjeksiyon yapmayınız .	• Psoriasis etkilenmiş bölgelere enjekte etmeyin
-----	---	---	--

○ Error-typology chart of the MT output:

Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use				0	0
Terminology	Incorrect term use	1	2		3	7
Accuracy	Mistranslation				0	0
Accuracy	Overtranslation				0	0
Accuracy	Undertranslation				0	0
Accuracy	Addition				0	0
Accuracy	Omission				0	0
Accuracy	Untranslated		2		2	6
Linguistic Conve	Grammar	1			1	1
Linguistic Conve	Spelling				0	0
Syntax	Syntactical errors				0	0
Style & Clarity	Unnatural Phrasing	1		1	2	6
Style & Clarity	Poor Readability			1	1	5
Custom	Layout, character & other				0	0
		Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
		3	4	2	9	25

Figure 28: Error-typology Chart (GPT-4 – zero-shot)

In Segment 3, the term “pre-filled pen” was translated as “dolu kalem”. These types of pharmaceutical products, namely pre-filled pens that contain solution for injection, are not referred to as “dolu kalem” in terms of target language conventions, i.e: Turkish language conventions in pharmaceutical context. They are commonly referred to as “enjeksiyon kalemi” or “kullanıma hazır kalem”.

In Segment 6, the collocation “special disposal container” was rendered in reference text as “özel atık kabı”. In the TT (reference), for the term “container”, the term “kap” was preferred. In the MT output, the term “konteyner” was used. Although not a crucial deviation, this is slightly an inaccurate denotation. “Konteyner” is a borrowing term that signifies recycling cans that are used in public and large in size. However, in pharmaceutical settings,

needles are disposed via much tinier containers which can be deemed a “kap” rather than a “konteyner” in terms of Turkish language conventions.

Segment 12 has an unnatural usage error which lies in the collocation “göbek düğmeniz” of the MT output. The term “belly button” is the equivalent of “göbek deliği”. The latter is a fixed, common ground usage that is accepted by both laymen and professionals. A Turkish speaker will not refer this to as “göbek düğmesi”. Although the term “button”'s denotative meaning is “düğme”, Turkish language refers to belly button area as “göbek deliği”. The term “delik” is the equivalent for the term “hole” in English. Therefore, to translate literally, Turkish speakers refer to “belly button” as “the hole of the belly”. This has been marked as a crucial unnatural phrasing error when performing TQE error analysis.

To continue from the same segment -Segment 12-, just as Google Translate’s MT output, the area to perform injection on is not clearly defined. There can be observed a serious ambiguity in the following sequence: “göbeğiniz (karın), göbek düğmenizden (navel) en az 5 cm uzaklıkta”. It doesn’t offer clear guidance on exactly where to perform injection on the abdomen area and what the 5 cm distance from belly button is about, distance to where and so forth. This segment is one of the most crucial segments of the entire extract because it gives guidance on an injection area which directly impacts patient safety and effective use of the product mentioned.

Same as Google Translate’s MT output, Segment 17 has the incorrect term use error of the term “red”, which should have been referred to as “kızarmış” rather than “kırmızı”. This can lead to ambiguities because people who have congenital red skin spots, for example, may perceive the instruction as to avoid all types of red -including congenital colour- and may miss the fact that it is about the redness that has occurred due to an incident, such as the redness due to burnt or the redness due to receiving damage, etc.

Lastly, in Segment 18, the term “psoriasis” has both been left untranslated in the MT output, and also there is a grammatical issue: “Psoriasis etkilenmiş bölgelere enjekte etmeyin”. The scientific term for the disease should have been either Turkicized or translated as “sedef hastalığı”. As for the grammar issue, it lies in missing ablative case in Turkish: “-den/-dan” which gives the meaning “by (something)”. The phrase which has just been

mentioned needs, despite being left untranslated, an ablative case, i.e: “Psoriasis-ten etkilenmiş”. Nevertheless, instead of such rendering, it should have been rendered as (such as in the reference text) “sedef hastalığından etkilenmiş”.

5.3.5.3 GPT-4 – RBP

Table 40: ST-3, Reference Text for ST-3, and MT Output (GPT-4 - RBP)

Segment No.	Source Text:	Reference Text:	MT Output:
1.	STEP 2	ADIM 2	ADIM 2
2.	Place the following items on a clean, flat surface:	Aşağıdaki malzemeleri temiz ve düz bir yüzey üzerine yerleştiriniz:	Temiz, düz bir yüzey üzerine aşağıdaki eşyaları yerleştirin:
3.	• 1 pre-filled pen	• 1 adet kullanıma hazır kalem	• 1 adet dolu enjektör kalem
4.	• 1 alcohol pad (not included in the carton)	• 1 adet alkollü ped (kutuya dahil değildir)	• 1 adet alkol pedi (kutuda bulunmamaktadır)
5.	• 1 cotton ball or gauze pad (not included in the carton)	• 1 adet pamuk yumağı veya gazlı bez (kutuya dahil değildir)	• 1 adet pamuk topu veya gazlı bez (kutuda bulunmamaktadır)
6.	• special disposal container (not included in the carton)	• Özel atık kabı (kutuya dahil değildir)	• Özel atık kutusu (kutuda bulunmamaktadır)
7.	Wash and dry your hands.	Ellerinizi yıkayıp kurulayınız.	Ellerinizi yıkayın ve kurulayın.
8.	STEP 3	ADIM 3	ADIM 3
9.	Choose from these 3 areas to inject:	Enjeksiyon için şu 3 bölgeden birini seçiniz:	Enjeksiyon yapılacak üç alan arasından birini seçin:
10.	• front of left thigh	• sol uyluğun önü	• sol uyluğun ön kısmı
11.	• front of right thigh	• sağ uyluğun önü	• sağ uyluğun ön kısmı

Table 40 - continue

12.	• your belly (abdomen) at least 5 cm from your belly button (navel)	• göbek deliğinin en az 5 cm uzağında olmak üzere göbek bölgesi (karın)	• göbeğinizin (karın) en az 5 cm uzağında
13.	Before the injection, wipe where you will inject in a circular motion with an alcohol pad.	Enjeksiyondan önce, enjeksiyon yapılacak yeri bir alkollü ped ile dairesel hareketlerle siliniz.	Enjeksiyon yapmadan önce, enjeksiyon yapılacak yeri alkol pedi ile dairesel hareketlerle silin.
14.	• Do not touch or blow on the injection site after it is cleaned.	• Temizledikten sonra enjeksiyon bölgesine dokunmayınız ya da üfleme yapmayınız.	• Temizlendikten sonra enjeksiyon yerine dokunmayın veya üfleme yapmayın.
15.	Allow the skin to dry before injecting	Enjeksiyondan önce derinizin kurumasını bekleyiniz.	Enjekte etmeden önce cildin kurumasını bekleyin.
16.	• Do not inject through clothes	• Kıyafetlerin üstünden enjeksiyon yapmayınız .	• Kıyafetlerin üzerinden enjekte etmeyin.
17.	• Do not inject into skin that is sore, bruised, red, hard, scarred, or has stretch marks	• Ağrıyan, morarmış, kızarmış, sertleşmiş, yaralı veya çatlak izleri olan deri üzerine enjeksiyon yapmayınız .	• Ağrılı, morarmış, kızarmış, sert, yara izli veya çatlaklı ciltlere enjeksiyon yapmayın.
18.	• Do not inject into areas affected by psoriasis	• Sedef hastalığından etkilenmiş bölgelere enjeksiyon yapmayınız .	• Sedef hastalığı olan bölgelere enjeksiyon yapmayın.

o Error-typology chart of the MT output:

Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use				0	0
Terminology	Incorrect term use				0	0
Accuracy	Mistranslation				0	0
Accuracy	Overtranslation				0	0
Accuracy	Undertranslation				0	0
Accuracy	Addition				0	0
Accuracy	Omission			1	1	5
Accuracy	Untranslated				0	0
Linguistic Conve	Grammar				0	0
Linguistic Conve	Spelling				0	0
Syntax	Syntactical errors				0	0
Style & Clarity	Unnatural Phrasing	1			1	1
Style & Clarity	Poor Readability			1	1	5
Custom	Layout, character & other		1		1	3
		Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
		1	1	2	4	14

Figure 29: Error-typology Chart (GPT-4 – RBP)

This output is remarkably similar to its raw-prompted output. They have common errors in general. However, one different and crucial error of this RBP output is that in Segment 12, such as in Google Translate’s MT output, “belly button” was translated as “göbeğinizden” which translates as “from your belly”, although in ST belly button area is clearly described. “Göbek deliği” being omitted has been evaluated as a critical omission error because it directly impacts patient safety.

The raw-prompted MT output has 25 TQE penalty score whereas RBP output has 14 TQE penalty score. Although drastically declined, this role-based prompted output, with its errors, shows us that the RBP method hasn’t necessarily increased translation quality as both outputs still have common errors. That is to say, the majority of the errors in the raw-prompted output hasn’t been eliminated, thus, the quality of the outputs remained similar.

5.3.5.4 Gemini-2.0-flash – zero-shot

Table 41: ST-3, Reference Text for ST-3, and MT Output (Gemini-2.0-flash – zero-shot)

Segment No.	Source Text:	Reference Text:	MT Output:
1.	STEP 2	ADIM 2	ADIM 2
2.	Place the following items on a clean, flat surface:	Aşağıdaki malzemeleri temiz ve düz bir yüzey üzerine yerleştiriniz:	Aşağıdaki malzemeleri temiz, düz bir yüzeye yerleştirin:
3.	• 1 pre-filled pen	• 1 adet kullanıma hazır kalem	• 1 adet önceden doldurulmuş kalem
4.	• 1 alcohol pad (not included in the carton)	• 1 adet alkollü ped (kutuya dahil değildir)	• 1 adet alkollü mendil (kartona dahil değildir)
5.	• 1 cotton ball or gauze pad (not included in the carton)	• 1 adet pamuk yumağı veya gazlı bez (kutuya dahil değildir)	• 1 adet pamuklu çubuk veya gazlı bez (kartona dahil değildir)
6.	• special disposal container (not included in the carton)	• Özel atık kabı (kutuya dahil değildir)	• özel atık kutusu (kartona dahil değildir)
7.	Wash and dry your hands.	Ellerinizi yıkayıp kurulayınız.	Ellerinizi yıkayın ve kurulayın.
8.	STEP 3	ADIM 3	ADIM 3
9.	Choose from these 3 areas to inject:	Enjeksiyon için şu 3 bölgeden birini seçiniz:	Enjeksiyon için bu 3 bölgeden birini seçin:
10.	• front of left thigh	• sol uyluğun önü	• sol uyluğun ön kısmı
11.	• front of right thigh	• sağ uyluğun önü	• sağ uyluğun ön kısmı
12.	• your belly (abdomen) at least 5 cm from your belly button (navel)	• göbek deliğinin en az 5 cm uzağında olmak üzere göbek bölgesi (karın)	• göbek deliğinizden (navel) en az 5 cm uzakta karnınız (abdomen)

Table 41 - continue

13.	Before the injection, wipe where you will inject in a	Enjeksiyondan önce, enjeksiyon yapılacak yeri	Enjeksiyondan önce, enjeksiyon yapacağınız yeri alkollü mendille dairesel hareketlerle silin.
	circular motion with an alcohol pad.	bir alkollü ped ile dairesel hareketlerle siliniz.	
14.	• Do not touch or blow on the injection site after it is cleaned.	• Temizledikten sonra enjeksiyon bölgesine dokunmayınız ya da üfleme yapmayınız.	• Temizlendikten sonra enjeksiyon bölgesine dokunmayın veya üfleme yapmayın.
15.	Allow the skin to dry before injecting	Enjeksiyondan önce derinizin kurumasını bekleyiniz.	Enjeksiyondan önce cildin kurumasına izin verin.
16.	• Do not inject through clothes	• Kıyafetlerin üstünden enjeksiyon yapmayınız .	• Giysilerin üzerinden enjeksiyon yapmayın.
17.	• Do not inject into skin that is sore, bruised, red, hard, scarred, or has stretch marks	• Ağrıyan, morarmış, kızarmış, sertleşmiş, yaralı veya çatlak izleri olan deri üzerine enjeksiyon yapmayınız .	• Ağrılı, morarmış, kızarıklık, sert, yara izi olan veya çatlakları olan cilde enjeksiyon yapmayın.
18.	• Do not inject into areas affected by psoriasis	• Sedef hastalığından etkilenmiş bölgelere enjeksiyon yapmayınız .	• Sedef hastalığından etkilenen bölgelere enjeksiyon yapmayın.

o Error-typology chart of the MT output:

Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use				0	0
Terminology	Incorrect term use		3	1	4	14
Accuracy	Mistranslation				0	0
Accuracy	Overtranslation				0	0
Accuracy	Undertranslation				0	0
Accuracy	Addition				0	0
Accuracy	Omission				0	0
Accuracy	Untranslated		1		1	3
Linguistic Conve	Grammar	1			1	1
Linguistic Conve	Spelling				0	0
Syntax	Syntactical errors				0	0
Style & Clarity	Unnatural Phrasing		1		1	3
Style & Clarity	Poor Readability				0	0
Custom	Layout, character & other				0	0
		Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
		1	5	1	7	21

Figure 30: Error-typology Chart (Gemini-2.0-flash – zero-shot)

In Segments 3 and 13, the alcohol pad, that is referred as one of the items to be placed on a surface and through which the injection site is to be wiped, has been rendered as “alkollü mendil” which in fact means “an alcohol-based tissue” which can cause misleading guidance. This has been marked as an incorrect term error with critical severity because the term “pad” was rendered as “mendil (tissue)” instead of “ped (pad)”.

In Segment 5, the term “cotton ball” was translated in MT output as “pamuklu çubuk” which translates as “cotton bud”. This has been marked as a critical mistranslation error. The same error with the term “carton” as in Google Translate’s MT output persists here. The term “carton” was rendered as “karton” rather than “kutu”.

5.3.5.5 Gemini-2.0-flash – RBP

Table 42: ST-3, Reference Text for ST-3, and MT Output (Gemini-2.0-flash – RBP)

Segment No.	Source Text:	Reference Text:	MT Output:
1.	STEP 2	ADIM 2	ADIM 2

Table 42 - continue

2.	Place the following items on a clean, flat surface:	Aşağıdaki malzemeleri temiz ve düz bir yüzey üzerine yerleştiriniz:	Aşağıdaki malzemeleri temiz ve düz bir yüzeye yerleştiriniz:
3.	• 1 pre-filled pen	• 1 adet kullanıma hazır kalem	• 1 adet önceden doldurulmuş kalem
4.	• 1 alcohol pad (not included in the carton)	• 1 adet alkollü ped (kutuya dahil değildir)	• 1 adet alkollü ped (kartona dahil değildir)
5.	• 1 cotton ball or gauze pad (not included in the carton)	• 1 adet pamuk yumağı veya gazlı bez (kutuya dahil değildir)	• 1 adet pamuklu çubuk veya gazlı bez (kartona dahil değildir)
6.	• special disposal container (not included in the carton)	• Özel atık kabı (kutuya dahil değildir)	• özel tıbbi atık kutusu (kartona dahil değildir)
7.	Wash and dry your hands.	Ellerinizi yıkayıp kurulayınız.	Ellerinizi yıkayınız ve kurulayınız.
8.	STEP 3	ADIM 3	ADIM 3
9.	Choose from these 3 areas to inject:	Enjeksiyon için şu 3 bölgeden birini seçiniz:	Enjeksiyon için aşağıdaki 3 bölgeden birini seçiniz:
10.	• front of left thigh	• sol uyluğun önü	• sol uyluğun ön yüzü
11.	• front of right thigh	• sağ uyluğun önü	• sağ uyluğun ön yüzü
12.	• your belly (abdomen) at least 5 cm from your belly button (navel)	• göbek deliğinin en az 5 cm uzağında olmak üzere göbek bölgesi (karın)	• göbek deliğinizden (umbilikus) en az 5 cm uzaklıkta karın (abdomen) bölgeniz
13.	Before the injection, wipe where you will inject in a circular motion with an alcohol pad.	Enjeksiyondan önce, enjeksiyon yapılacak yeri bir alkollü ped ile dairesel hareketlerle siliniz.	Enjeksiyondan önce, enjeksiyon yapacağınız bölgeyi alkollü ped ile dairesel hareketlerle siliniz.

Table 42 - continue

14.	• Do not touch or blow on the injection site after it is cleaned.	• Temizledikten sonra enjeksiyon bölgesine dokunmayınız ya da üfleme yapmayınız.	• Temizlendikten sonra enjeksiyon bölgesine dokunmayınız veya üfleme yapmayınız.
15.	Allow the skin to dry before injecting	Enjeksiyondan önce derinizin kurumasını bekleyiniz.	Enjeksiyondan önce cildin kurumasına izin veriniz.
16.	• Do not inject through clothes	• Kıyafetlerin üstünden enjeksiyon yapmayınız .	• Giysilerinizin üzerinden enjeksiyon yapmayınız.
17.	• Do not inject into skin that is sore, bruised, red, hard, scarred, or has stretch marks	• Ağrılan, morarmış, kızarmış, sertleşmiş, yaralı veya çatlak izleri olan deri üzerine enjeksiyon yapmayınız .	• Ağrılı, morarmış, kızarıklık, sertleşmiş, skar dokusu olan veya çatlakları bulunan cilde enjeksiyon yapmayınız.
18.	• Do not inject into areas affected by psoriasis	• Sedef hastalığından etkilenmiş bölgelere enjeksiyon yapmayınız .	• Psoriazisten etkilenmiş bölgelere enjeksiyon yapmayınız.

o Error-typology chart of the MT output:

Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use				0	0
Terminology	Incorrect term use		3		3	9
Accuracy	Mistranslation		1	1	2	8
Accuracy	Overtranslation			1	1	5
Accuracy	Undertranslation				0	0
Accuracy	Addition	1			1	1
Accuracy	Omission				0	0
Accuracy	Untranslated		1		1	3
Linguistic Conve	Grammar				0	0
Linguistic Conve	Spelling				0	0
Syntax	Syntactical errors				0	0
Style & Clarity	Unnatural Phrasing		1		1	3
Style & Clarity	Poor Readability				0	0
Custom	Layout, character & other				0	0
		Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
		1	6	2	9	29

Figure 31: Error-typology Chart (Gemini-2.0-flash – RBP)

This RBP output has a higher TQE penalty score (29) than its raw-prompted output (21). In Segment 5, “cotton ball” was translated as “pamuklu çubuk” which is an incorrect term usage as the latter literally translates as “cotton bud”. In Segment 17, the term “scarred” being rendered as “skar dokusu olan” is an error of overtranslation, plus, this will not signify a solid component in lay readers’ minds.

5.3.5.6 DeepSeek - R1 – zero-shot

Table 43: ST-3, Reference Text for ST-3, and MT Output (DeepSeek – R1 – zero-shot)

Segment No.	Source Text:	Reference Text:	MT Output:
1.	STEP 2	ADIM 2	ADIM 2
2.	Place the following items on a clean, flat surface:	Aşağıdaki malzemeleri temiz ve düz bir yüzey üzerine yerleştiriniz:	Aşağıdaki malzemeleri temiz ve düz bir yüzeye yerleştirin:
3.	• 1 pre-filled pen	• 1 adet kullanıma hazır kalem	• 1 adet önceden doldurulmuş kalem
4.	• 1 alcohol pad (not included in the carton)	• 1 adet alkollü ped (kutuya dahil değildir)	• 1 adet alkollü ped (kutuya dahil değildir)
5.	• 1 cotton ball or gauze pad (not included in the carton)	• 1 adet pamuk yumağı veya gazlı bez (kutuya dahil değildir)	• 1 adet pamuk topu veya gazlı bez (kutuya dahil değildir)
6.	• special disposal container (not included in the carton)	• Özel atık kabı (kutuya dahil değildir)	• özel atık kabı (kutuya dahil değildir)
7.	Wash and dry your hands.	Ellerinizi yıkayıp kurulayınız.	Ellerinizi yıkayın ve kurulayın.
8.	STEP 3	ADIM 3	ADIM 3
9.	Choose from these 3 areas to inject:	Enjeksiyon için şu 3 bölgeden birini seçiniz:	Enjeksiyon için şu 3 bölgeden birini seçin:

Table 43 - continue

10.	• front of left thigh	· sol uyluğun önü	• Sol uyluğun ön kısmı
11.	• front of right thigh	· sağ uyluğun önü	• Sağ uyluğun ön kısmı
12.	• your belly (abdomen) at least 5 cm from your belly button (navel)	· göbek deliğinin en az 5 cm uzağında olmak üzere göbek bölgesi (karın)	• Karnınız (abdomen), göbek deliğinizden en az 5 cm uzakta
13.	Before the injection, wipe where you will inject in a circular motion with an alcohol pad.	Enjeksiyondan önce, enjeksiyon yapılacak yeri bir alkollü ped ile dairesel hareketlerle siliniz.	Enjeksiyondan önce, alkollü pedle enjeksiyon yapacağınız bölgeyi dairesel hareketlerle silin.
14.	• Do not touch or blow on the injection site after it is cleaned.	· Temizledikten sonra enjeksiyon bölgesine dokunmayınız ya da üfleme yapmayınız.	• Temizlenen bölgeye dokunmayın veya üfleme yapmayın.
15.	Allow the skin to dry before injecting	Enjeksiyondan önce derinizin kurumasını bekleyiniz.	Enjeksiyon yapmadan önce cildin kurumasını bekleyin.
16.	• Do not inject through clothes	· Kıyafetlerin üstünden enjeksiyon yapmayınız .	• Giysilerin üzerinden enjeksiyon yapmayın.
17.	• Do not inject into skin that is sore, bruised, red, hard, scarred, or has stretch marks	· Ağrıyan, morarmış, kızarmış, sertleşmiş, yaralı veya çatlak izleri olan deri üzerine enjeksiyon yapmayınız .	• Ağrılı, çürük, kızarıklık, sert, yara izli veya çatlak olan cilt bölgelerine enjeksiyon yapmayın.
18.	• Do not inject into areas affected by psoriasis	· Sedef hastalığından etkilenmiş bölgelere enjeksiyon yapmayınız .	• Sedef hastalığından etkilenen bölgelere enjeksiyon yapmayın.

o Error-typology chart of the MT output:

Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use				0	0
Terminology	Incorrect term use			1	1	5
Accuracy	Mistranslation				0	0
Accuracy	Overtranslation				0	0
Accuracy	Undertranslation				0	0
Accuracy	Addition				0	0
Accuracy	Omission				0	0
Accuracy	Untranslated				0	0
Linguistic Conve	Grammar				0	0
Linguistic Conve	Spelling				0	0
Syntax	Syntactical errors	1			1	1
Style & Clarity	Unnatural Phrasing		1		1	3
Style & Clarity	Poor Readability				0	0
Custom	Layout, character & other				0	0
		Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
		1	1	1	3	9

Figure 32: Error-typology Chart (DeepSeek – R1 – zero-shot)

In Segment 17, the term “bruised” was translated as “çürük” which translates literally as “necrotized” or “decayed”, whereas the correct usage would be “morarmış”.

5.3.5.7 DeepSeek - R1 – RBP

Table 44: ST-3, Reference Text for ST-3, and MT Output (DeepSeek – R1 – RBP)

Segment No.	Source Text:	Reference Text:	MT Output:
1.	STEP 2	ADIM 2	ADIM 2
2.	Place the following items on a clean, flat surface:	Aşağıdaki malzemeleri temiz ve düz bir yüzey üzerine yerleştiriniz:	Aşağıdaki malzemeleri temiz ve düz bir yüzeye yerleştirin:
3.	• 1 pre-filled pen	• 1 adet kullanıma hazır kalem	• 1 adet önceden doldurulmuş kalem
4.	• 1 alcohol pad (not included in the carton)	• 1 adet alkollü ped (kutuya dahil değildir)	• 1 adet alkollü temizleme pedi (kutu içinde yer almaz)

Table 44 - continue

5.	• 1 cotton ball or gauze pad (not included in the carton)	• 1 adet pamuk yumağı veya gazlı bez (kutuya dahil değildir)	• 1 adet pamuk topu veya gazlı bez (kutu içinde yer almaz)
6.	• special disposal container (not included in the carton)	• Özel atık kabı (kutuya dahil değildir)	• özel atık kabı (kutu içinde yer almaz)
7.	Wash and dry your hands.	Ellerinizi yıkayıp kurulayınız.	Ellerinizi yıkayın ve kurulayın.
8.	STEP 3	ADIM 3	ADIM 3
9.	Choose from these 3 areas to inject:	Enjeksiyon için şu 3 bölgeden birini seçiniz:	Enjeksiyon için aşağıdaki 3 bölgeden birini seçin:
10.	• front of left thigh	• sol uyluğun önü	• sol uyluğun ön kısmı
11.	• front of right thigh	• sağ uyluğun önü	• sağ uyluğun ön kısmı
12.	• your belly (abdomen) at least 5 cm from your belly button (navel)	• göbek deliğinin en az 5 cm uzağında olmak üzere göbek bölgesi (karın)	• karın bölgesi (abdomen), göbek deliğinizden (göbek) en az 5 cm uzakta
13.	Before the injection, wipe where you will inject in a circular motion with an alcohol pad.	Enjeksiyondan önce, enjeksiyon yapılacak yeri bir alkollü ped ile dairesel hareketlerle siliniz.	Enjeksiyon öncesinde, enjeksiyon yapacağınız bölgeyi alkollü ped ile dairesel hareketlerle silin.
14.	• Do not touch or blow on the injection site after it is cleaned.	• Temizledikten sonra enjeksiyon bölgesine dokunmayınız ya da üfleme yapınız.	• Temizlenen bölgeye dokunmayın veya üfleme yapmayın.
15.	Allow the skin to dry before injecting	Enjeksiyondan önce derinizin kurumasını bekleyiniz.	Enjeksiyon öncesinde cildin kurumasını bekleyin.
16.	• Do not inject through clothes	• Kıyafetlerin üstünden enjeksiyon yapmayınız .	• Giysilerin üzerinden enjeksiyon yapmayın.
17.	• Do not inject into skin that is sore, bruised, red, hard, scarred, or has stretch marks	• Ağrıyan, morarmış, kızarmış, sertleşmiş, yaralı veya çatlak izleri olan deri	• Ağrılı, morarmış, kızarık, sertleşmiş, skarlı veya çatlak izleri olan cilde enjeksiyon yapmayın.

Table 44 - continue

		üzerine enjeksiyon yapmayınız.	
18.	• Do not inject into areas affected by psoriasis	• Sedef hastalığından etkilenmiş bölgelere enjeksiyon yapmayınız.	• Sedef hastalığından (psoriasis) etkilenen bölgelere enjeksiyon yapmayın.

o Error-typology chart of the MT output:

Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use				0	0
Terminology	Incorrect term use				0	0
Accuracy	Mistranslation				0	0
Accuracy	Overtranslation			1	1	5
Accuracy	Undertranslation				0	0
Accuracy	Addition				0	0
Accuracy	Omission				0	0
Accuracy	Untranslated				0	0
Linguistic Conve	Grammar				0	0
Linguistic Conve	Spelling				0	0
Syntax	Syntactical errors	1			1	1
Style & Clarity	Unnatural Phrasing		1		1	3
Style & Clarity	Poor Readability				0	0
Custom	Layout, character & other				0	0
		Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
		1	1	1	3	9

Figure 33: Error-typology Chart (DeepSeek – R1 – RBP)

This RBP output has the exact same TQE penalty score as its raw-prompted version, which is 9. There hasn't been a remarkable improvement noted in terms of the quality of the MT output, compared to raw-prompted output. Besides TQE penalty score, when it comes to BLEU metric, the overall BLEU of this output is 27.05 whereas the raw-prompted output has 39.61 points and is the highest among all the outputs. RBP output's BLEU rank is 6 out of 9. In this sense, the RBP output does not demonstrate a sharp rise.

5.3.5.8 Claude-3.5-Sonnet – zero-shot

Table 45: ST-3, Reference Text for ST-3, and MT Output (Claude-3.5-Sonnet – zero-shot)

Segment No.	Source Text:	Reference Text:	MT Output:
1.	STEP 2	ADIM 2	ADIM 2
2.	Place the following items on a clean, flat surface:	Aşağıdaki malzemeleri temiz ve düz bir yüzey üzerine yerleştiriniz:	Aşağıdaki malzemeleri temiz, düz bir yüzeye yerleştirin:
3.	• 1 pre-filled pen	• 1 adet kullanıma hazır kalem	1 önceden doldurulmuş kalem
4.	• 1 alcohol pad (not included in the carton)	• 1 adet alkollü ped (kutuya dahil değildir)	1 alkollü ped (karton kutuda bulunmamaktadır)
5.	• 1 cotton ball or gauze pad (not included in the carton)	• 1 adet pamuk yumağı veya gazlı bez (kutuya dahil değildir)	1 pamuk topu veya gazlı bez (karton kutuda bulunmamaktadır)
6.	• special disposal container (not included in the carton)	• Özel atık kabı (kutuya dahil değildir)	özel atık kabı (karton kutuda bulunmamaktadır)
7.	Wash and dry your hands.	Ellerinizi yıkayıp kurulayınız.	Ellerinizi yıkayıp kurulayın.
8.	STEP 3	ADIM 3	ADIM 3
9.	Choose from these 3 areas to inject:	Enjeksiyon için şu 3 bölgeden birini seçiniz:	Enjeksiyon için şu 3 bölgeden birini seçin:
10.	• front of left thigh	• sol uyluğun önü	sol uyluğun ön kısmı
11.	• front of right thigh	• sağ uyluğun önü	sağ uyluğun ön kısmı
12.	• your belly (abdomen) at least 5 cm from your belly button (navel)	• göbek deliğinin en az 5 cm uzağında olmak üzere göbek bölgesi (karın)	göbek deliğinizden (naval) en az 5 cm uzaklıkta karnınız (abdomen)

Table 45 - continue

13.	Before the injection, wipe where you will inject in a circular motion with an alcohol pad.	Enjeksiyondan önce, enjeksiyon yapılacak yeri bir alkollü ped ile dairesel hareketlerle siliniz.	Enjeksiyondan önce, enjeksiyon yapacağınız yeri dairesel hareketlerle alkollü pedle silin.
14.	• Do not touch or blow on the injection site after it is cleaned.	• Temizledikten sonra enjeksiyon bölgesine dokunmayınız ya da üfleme yapınız.	Temizledikten sonra enjeksiyon bölgesine dokunmayın veya üfleme yapmayın.
15.	Allow the skin to dry before injecting	Enjeksiyondan önce derinizin kurumasını bekleyiniz.	Enjeksiyon yapmadan önce cildin kurumasını bekleyin
16.	• Do not inject through clothes	• Kıyafetlerin üstünden enjeksiyon yapmayınız .	Kıyafetlerin üzerinden enjeksiyon yapmayın
17.	• Do not inject into skin that is sore, bruised, red, hard, scarred, or has stretch marks	• Ağrıyan, morarmış, kızarmış, sertleşmiş, yaralı veya çatlak izleri olan deri üzerine enjeksiyon yapmayınız .	Ağrılı, morarmış, kızarmış, sertleşmiş, yara izli veya çatlakları olan cilde enjeksiyon yapmayın
18.	• Do not inject into areas affected by psoriasis	• Sedef hastalığından etkilenmiş bölgelere enjeksiyon yapmayınız .	Sedef hastalığından etkilenen bölgelere enjeksiyon yapmayın

- o Error-typology chart of the MT output:

Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use				0	0
Terminology	Incorrect term use				0	0
Accuracy	Mistranslation				0	0
Accuracy	Overtranslation				0	0
Accuracy	Undertranslation				0	0
Accuracy	Addition		3		3	9
Accuracy	Omission				0	0
Accuracy	Untranslated				0	0
Linguistic Conve	Grammar				0	0
Linguistic Conve	Spelling				0	0
Syntax	Syntactical errors				0	0
Style & Clarity	Unnatural Phrasing		1		1	3
Style & Clarity	Poor Readability	1			1	1
Custom	Layout, character & other				0	0
		Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
		1	4	0	5	13

Figure 34: Error-typology Chart (Claude-3.5-Sonnet – zero-shot)

5.3.5.9 Claude-3.5-Sonnet – RBP

Table 46: ST-3, Reference Text for ST-3, and MT Output (Claude-3.5-Sonnet – RBP)

Segment No.	Source Text:	Reference Text:	MT Output:
1.	STEP 2	ADIM 2	ADIM 2
2.	Place the following items on a clean, flat surface:	Aşağıdaki malzemeleri temiz ve düz bir yüzey üzerine yerleştiriniz:	Aşağıdaki malzemeleri temiz, düz bir yüzeye yerleştirin:
3.	• 1 pre-filled pen	• 1 adet kullanıma hazır kalem	1 kullanıma hazır kalem
4.	• 1 alcohol pad (not included in the carton)	• 1 adet alkollü ped (kutuya dahil değildir)	1 alkollü ped (kutu içinde bulunmaz)
5.	• 1 cotton ball or gauze pad (not included in the carton)	• 1 adet pamuk yumağı veya gazlı bez (kutuya dahil değildir)	1 pamuk topu veya gazlı bez (kutu içinde bulunmaz)
6.	• special disposal container (not included in the carton)	• Özel atık kabı (kutuya dahil değildir)	özel atık kabı (karton kutuda bulunmamaktadır)

Table 46 - continue

7.	Wash and dry your hands.	Ellerinizi yıkayıp kurulayınız.	Ellerinizi yıkayıp kurulayın.
8.	STEP 3	ADIM 3	ADIM 3
9.	Choose from these 3 areas to inject:	Enjeksiyon için şu 3 bölgeden birini seçiniz:	Enjeksiyon için şu 3 bölgeden birini seçin:
10.	• front of left thigh	• sol uyluğun önü	sol uyluğun ön kısmı
11.	• front of right thigh	• sağ uyluğun önü	sağ uyluğun ön kısmı
12.	• your belly (abdomen) at least 5 cm from your belly button (navel)	• göbek deliğinin en az 5 cm uzağında olmak üzere göbek bölgesi (karın)	göbek deliğinizden en az 5 cm uzaklıkta karnınız (abdomen)
13.	Before the injection, wipe where you will inject in a circular motion with an alcohol pad.	Enjeksiyondan önce, enjeksiyon yapılacak yeri bir alkollü ped ile dairesel hareketlerle siliniz.	Enjeksiyon öncesinde, enjeksiyon yapacağınız bölgeyi alkollü pedle dairesel hareketle silin.
14.	• Do not touch or blow on the injection site after it is cleaned.	• Temizledikten sonra enjeksiyon bölgesine dokunmayınız ya da üfleme yapmayınız.	Temizledikten sonra enjeksiyon bölgesine dokunmayın veya üzerine üfleme yapmayın.
15.	Allow the skin to dry before injecting	Enjeksiyondan önce derinizin kurumasını bekleyiniz.	Enjeksiyon yapmadan önce derinin kurumasını bekleyin
16.	• Do not inject through clothes	• Kıyafetlerin üstünden enjeksiyon yapmayınız .	Giysilerin üzerinden enjeksiyon yapmayın
17.	• Do not inject into skin that is sore, bruised, red, hard, scarred, or has stretch marks	• Ağrıyan, morarmış, kızarmış, sertleşmiş, yaralı veya çatlak izleri olan deri üzerine enjeksiyon yapmayınız .	Ağrılı, çürük, kızarıklık, sert, yara izli veya çatlakları olan cilde enjeksiyon yapmayın

Table 46 - continue

18.	• Do not inject into areas affected by psoriasis	· Sedef hastalığından etkilenmiş bölgelere enjeksiyon yapmayınız.	Sedef hastalığından etkilenen bölgelere enjeksiyon yapmayın
-----	---	--	---

o Error-typology chart of the MT output:

Quality Issue	Error Type	Minor	Major	Critical	Sum of Instances	Penalty Point
Terminology	Inconsistent use				0	0
Terminology	Incorrect term use			1	1	5
Accuracy	Mistranslation				0	0
Accuracy	Overtranslation				0	0
Accuracy	Undertranslation				0	0
Accuracy	Addition				0	0
Accuracy	Omission				0	0
Accuracy	Untranslated				0	0
Linguistic Conve	Grammar				0	0
Linguistic Conve	Spelling				0	0
Syntax	Syntactical errors				0	0
Style & Clarity	Unnatural Phrasing				0	0
Style & Clarity	Poor Readability	1			1	1
Custom	Layout, character & other				0	0
		Minor Total	Major Total	Critical Total	Total Error Instances	Final Penalty Score
		1	0	1	2	6

Figure 35: Error-typology Chart (Claude-3.5-Sonnet – RBP)

5.3.6 Overall Results Overview for ST-3’s TT Outputs

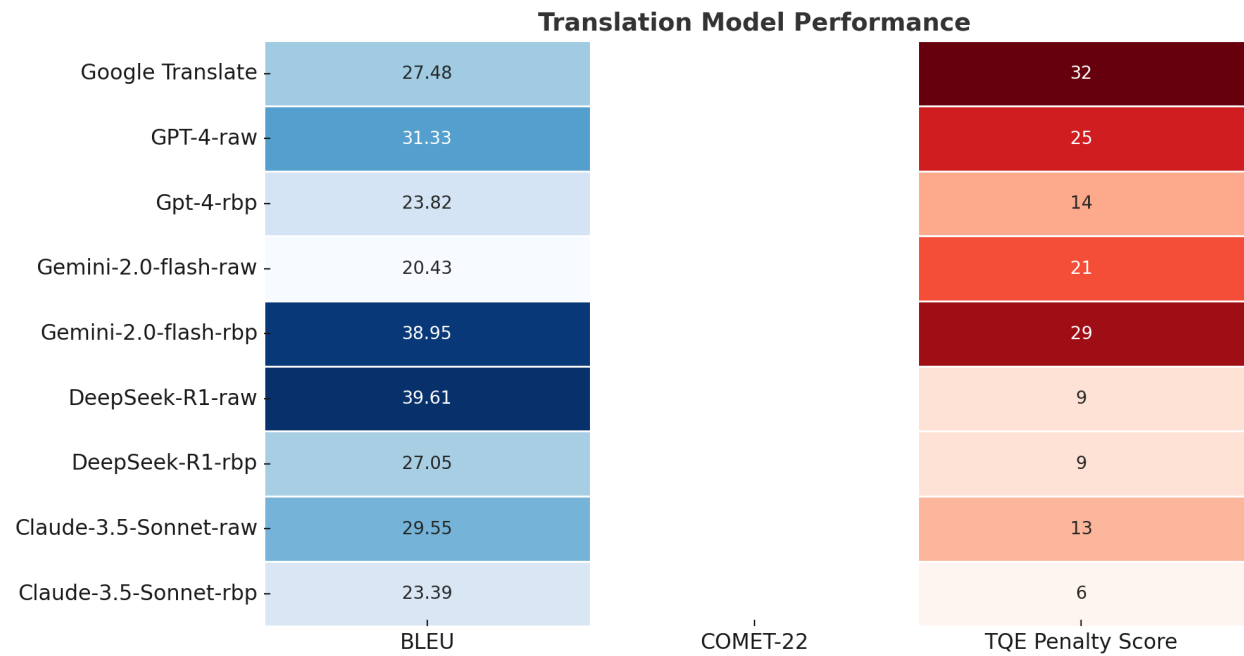


Figure 36: Heatmap Overview of Quantitative and Qualitative Metric Results of ST-3’s TT Outputs

6. DISCUSSION

6.1 Text-type and Function Analysis

Texts have different types depending on their functions and the interaction they establish with the intended audiences. During the course of planning how to structure a text, text producers choose one of the three fundamental communicative types of text: informative text type, expressive text type, and operative text type (cf. Reiss and Vermeer, 2014:137). Informative text type is the type of text where pieces of information are conveyed to the audience, expressive text type emphasizes aesthetic or artistic expression, and operative text type aims to persuade or elicit a response from the reader (ibid). With respect to pharmaceutical texts, particularly the ones under the scope of this study: PIL and SmPC documents, they are primarily informative (with PIL documents having secondarily operative elements as well which will be defined in the following paragraph).

The main purpose of a PIL document is to convey clear, factual medical information to patients including dosage instructions, side effects, injection sites, storage conditions, destruction guidelines, etc. In some cases, however, certain sections of a PIL document may also have operative elements. When, for instance, offering a caution to patients on following instructions carefully, or advising them to seek medical advice or to avoid certain behaviours, these sections aim to influence patient behaviour which aligns with the operative function mentioned by Reiss and Vermeer (2014).

The main function of an SmPC document is to convey technical, scientific and regulatory information about a medicinal product like its composition, pharmacodynamics, clinical trial data, contraindications, interactions etc. SmPC documents are structured and standardized texts that need to comply with frameworks of regulatory bodies such as EMA (European Medicines Agency) and FDA (Food and Drug Administration). An SmPC document aims to offer accurate information, with a formal, technical and clear style, which is intended to address to healthcare professionals rather than patients. In this sense, this is also categorized under informative text type. This, however, also raises an important

question: how can a text like this work just as well in another language and culture? At this point, Christiane Nord's functionalist model (2006) is worth mentioning, because it focuses on translating with a clear purpose and taking target audience into account, rather than solely changing words from one language to another.

Following Christiane Nord's functionalist model of translation, each of the three source texts (ST-1, ST-2, and ST-3) in this study serves a communicative purpose, and therefore requires well-designed translation strategies. According to Nord, translators determine their strategies shaped by the purpose and function, in alignment with the audience which the translation will need to address to (cf. Nord, 2006:131). In addition, Nord adds:

Every translation is intended to achieve a particular communicative purpose in the target audience, and if we analyse who the target audience will be and what they may need and expect, we might be better able to deliver a product that suits their needs and expectations. (Nord, 2006:133)

ST-1, which is an extract from an SmPC document for the pharmaceutical drug Opdivo, is a technical and informative text aimed at healthcare professionals. Its main goal is to offer precise clinical and pharmaceutical information, which means the TT needs to follow compliant medical terminology and regulatory standards.

ST-2 and ST-3, on the other hand, are both taken from PILs and are addressed to the general public, especially patients. These texts also aim to inform the reader, but they include instructions, warnings, or advice such as "Do not breast-feed while taking this medicine," which are intended to influence readers' actions. According to Nord, this is an appellative function, where the goal is to encourage the reader to act in a certain way (cf. Nord, 2006:138). Because these texts are directed at lay readers (namely non-experts) their translations must use plain, clear, and accessible language to help patient understanding.

To sum up, based on the discussions above, both Reiss and Nord help us understand that different texts have different purposes, and translators need to be aware of these when working with medical texts like PILs and SmPCs. While SmPCs are aimed at healthcare professionals and must be translated with technical accuracy, PILs are meant for lay readers and need to be easy to read and understand.

6.2 Domain-specific Oversight

The domain-specific characteristics of medical and pharmaceutical texts played an important role in shaping both the design and conduct of this study. Contrary to LGP texts, medical texts, especially SmPCs and PILs, contain a hybrid of LSP and LGP, that is to say they include high amount of terminological elements, thus this brings limitations that depend on the context. This factor poses a remarkable challenge for MT systems especially when the MT system is not specifically trained.

One major challenge observed during the course of the pilot experiments was the inconsistencies that occurred in specialized terminology. For instance, as can be seen in error annotations, terms like “randomised” were mistranslated as “rastgele” at times, instead of the established counterpart “randomize” or “randomize edilmiş” in the clinical language conventions of the Turkish language. This led to a lack of alignment with common-ground clinical Turkish terminology. Similarly, acronyms such as ECOG or expressions like “wild-type” were frequently erroneous, either left untranslated or inconsistently rendered, even in role-based prompted outputs. Another important insight is about register inconsistency. Some outputs from models occasionally failed to adhere to the formal tone required in professional medical texts.

Even when some BLEU scores were high, TQE analyses showed that the outputs lacked the precision and style expected in domain-specific communication, especially in expert-facing documents. In addition, as it can be seen from the research materials in the study, medical domain requires high risk-awareness. Errors such as in injection instructions are not just translation errors but also potentially hazardous risks since it can affect real practice of the reader. This crucial factor stresses out the limitations of relying solely on pre-trained general-purpose LLMs for domain-specific translation without post-editing or applying a preliminary training such as fine-tuning.

In this sense, the use of role-based prompting was a deliberate strategy to investigate if it could achieve domain adaptation. The findings, however, showed that even with domain-

specific instructions, the role-prompted LLMs did not necessarily outperform their zero-shot prompted models as it can be seen in metric results. This suggests that domain awareness and competence cannot be applied through prompting alone and that domain-specific training, for instance fine-tuning or retrieval-augmented generation (RAG) method, may become necessary for higher quality outputs in high-stakes domains such as medical/pharmaceutical texts.

To conclude, the results highlight the need for domain-aware oversight methods in addition to TQE analysis, especially when engaging AI systems into translation workflows in medical domain. To overcome these limitations, future work may be encouraged to focus on training LLMs with customized domain corpora and/or glossaries to try mitigating these limitations.

6.3 Would Human – Machine Collaboration be Possible?

The ongoing disputes in the era of translation technologies mainly revolve around the question whether machines would be an “enemy” or or a “replacement”. However, the evolution in translation industry has shown a shift towards human-machine collaboration which has been shaping both practice and theory.

Scholars like Tieber and Baumgarten (2024) describe this transformation as a “sociotechnical (r)evolution” (2024:379), highlighting the progressive integration of AI-driven translation tools into modern translation workflows (cf. Tieber and Baumgarten, 2024:380). While these tools need to be monitored closely, this also brings up concerns around autonomy, professional satisfaction, and labour conditions for translators (cf. Tieber and Baumgarten, 2024:384). Effective human-machine interaction requires more than technical functionality. “Translator-computer interaction (TCI)” (O’Brien, 2012:101, as cited in Tieber and Baumgarten, 2024:382) and translation technologies have been reshaping cognitive and ergonomic sides of translation activity (cf. Tieber and Baumgarten, 2024:382), alongside scholars recognizing and exploring the socioeconomic affect of this interaction in the digital global economy (cf. Tieber and Baumgarten, 2024:384).

Similar to these ideas, O'Brien (2024) suggests a shift from antagonistic dualisms - humans vs. machines- toward a human-centered augmented translation approach. She draws on Shneiderman's concept of Human-Centered AI (HCAI) (cf. Shneiderman, 2020a, 2020b, as cited in O'Brien, 2024:392) which advocates for systems that amplify human intelligence rather than replicate or replace it. In this approach, AI technologies work as "cognitive prosthetic" (cf. Stanney et al, 2015:330, as cited in O'Brien, 2024:396) that empowers cognitive power of translators' brain, enhances computational power of the computational agent and emphasizing a symbiotic relationship instead of a rivalry (cf. O'Brien, 2024:395). O'Brien repositions Haraway's (1985) cyborg metaphor which suggests that translators are already hybrid beings, combining human creativity with machine assistance (cf. O'Brien, 2024:392). This hybridity should be embraced rather than showing resistance.

Schumacher (2023:2) also advocate the "machine-assisted human translation" notion against "human-assisted MT". In her article, Schumacher argues that we should rethink the way we benefit from MT in translation activities. Instead of relying on the common approach where humans are merely actors to fix the errors made by MT, she proposes the opposite, that is machine-assisted human translation (cf. Schumacher, 2023:4). This means that the translator should remain in the centre of control in translation activities, using MT tools to amplify their own decisions and creativity (cf. Schumacher, 2023:2).

In addition to these perspectives, Öner and Bengi (2024) highlight the evolving role of human competence in a tech-driven language services industry. Despite the fears of obsolescence, they suggest that human competence: especially language skills, domain awareness (i.e: expertise of area of specialization), and technological literacy are irreplaceable (cf. Öner and Bengi, 2024:78). Their case study which focuses on ChatGPT's efficiency in generating content for user manuals shows the crucial role of human involvement in prompt design and quality assurance (cf. Öner and Bengi, 2024:78). They emphasize that, for achieving quality and effective content, careful monitoring is needed by skilled professionals, who possess the afore-mentioned skills (cf. Öner and Bengi, 2024:81). This stresses out the importance of human judgment in achieving reliable translation outcomes. Similar to what Öner and Bengi suggest, Reynolds and McDonnell (2021:4) note:

If we were to predict how a given passage of text would continue given that a human had written it, we would need to model the intentions of its writer and incorporate worldly knowledge about its referents.

Shanahan et al. (2023), in their work focusing on role-playing LLMs, argue that role-playing for dialogue agents (i.e: LLMs) is useful because it lets us recognize that these AI systems understanding human concepts (like beliefs, desires, goals, and emotions) without falsely thinking that they have human-like features (anthropomorphism) (cf. Shanahan et al., 2023:4). In other words, when we consider LLMs as actors playing roles, it helps us understand and monitor their behaviour, which also helps us remember that they're essentially inhuman (ibid.).

The authors, however, also stress that the metaphor of role-play has limitations. When we mention "role-play," we often imagine a human actuator whose language skills are a continuity of their cognitive capacities they have developed upon having interacted with the world (cf. Shanahan et al., 2023:1). Dialogue agents based on LLMs, however, don't exactly function this way. Instead of committing to one predefined character, a dialogue agent constantly generates and adapts to a variety of characters as the interaction progresses (cf. Shanahan et al, 2023:4). The authors add: "The dialogue agent is more like a performer in improvisational theatre than an actor in a conventional, scripted play" (ibid.). To sum up, on one hand role-play is an accessible way to discuss AI behavior, and on the other hand, Shanahan et al. suggest that it would even be more correct to describe an LLM as dynamically improvising agents across a set of potential characters, rather than performing a predefined single role.

The future of translation shouldn't be about humans versus machines, but should be about building a balance where both can function together. Instead of considering AI as a threat or necessarily a replacement, it makes more sense to view it as a tool that can assist translators to do their jobs more efficiently. It can be suggested that machines can handle variables such as speed and consistency well, however they still need human intervention for practices like understanding aspects like context, culture, and tone. When both parties

cooperate, the result to achieve can be more efficient than what either of the actuators can do alone.

It is also beneficial to remember that for such collaboration to work well, translation tools should be designed with the human user in mind. In order for this, it should be ensured that they are easy to use and actually support the translator's cognitive process. It's obvious that the prevalence and quality of machine agency will be increasingly ongoing (Öner Bulut, 2019:18) and this requires translator trainers alongside researchers to develop training methods to prepare the students for the future (cf. Öner Bulut, 2019:19). Training for future translators should also focus on learning how to collaborate with AI tools and how to approach them critically. It shouldn't be only about how to use them, but also how to guide them and check their output.

To sum up, human-machine collaboration in translation doesn't only mean using the latest tech but it's also about building a partnership. Machines may be efficient assistants when it comes to intelligence amplification, yet, it is the human translator who gives translated product its soul. In my opinion, the challenge that awaits us is to design ethical, and empowering frameworks that ensure this collaboration results in aligned outputs with the values of the profession.

6.4 Qualifications Expected from a Medical Translator

The sociology of a translation focuses on how a translated product emerges, how the constructs behind this product operate, what actuators perform and how they perform it (Demirel, 2014:403, as cited in Seçkin and Bogenç, 2019:112).¹⁷ "Translation Sociology", a concept that has emerged recently, has focused more on literary works and investigated the differences in social environments of authors and translators in target text production process, leaving out technical texts such as medical texts in the secondary place (cf. Taşdan Doğan, 2020:934-935).

¹⁷ Translation is my own.

Medical translators, however, also work in complex social environments especially because the translated product they produce can directly affect patient health and safety and result in crucial consequences. Including technical fields like medical translation in the frames that are being shaped around translation sociology would help us better understand medical translators' key role and responsibilities in these high-stakes texts. Recognizing the sociological aspects of medical translation hence requires a closer look into the competencies required of medical translators, as their work not only revolves around social contexts but also demands professional awareness and specialization.

Medical text translation is not a task to be adequately handled by any bilingual individual or any translator. A medical translator, of course, is expected to bring along certain qualifications with themselves, just like in any other specialized fields. Filatova (2017) lists and elaborates on 6 difference competences, according to ISO 17100:2015, a medical translator should possess:

- translation competence,
- linguistic and textual competence in the source and the target language,
- competence in research, information acquisition, and processing,
- cultural competence,
- technical competence,
- domain competence (cf. Filatova, 2017:67).

As Oğuz (2014:127) highlights, the translator must possess 'considerable scientific background and experience' and master terminology, acronyms, eponyms, and even the structural elements of Greek and Latin. Besides this perspective, a medical translator's perception of medical English and their writing skills should embody a set of registers and genres (cf. Karwacka, 2014:23) as translating medical texts is an activity that requires writing skills (ibid.).

This nature of medical translation shows us why human expertise becomes a must, especially when we take into account the increasing tendency in the use of LLMs as autonomous actuators in translation workflows. Despite their improving capabilities and ability to handle large linguistic data, LLMs, as can be seen in this experimental study, can

lack the domain-specific competence, contextual awareness, and pragmatic sensitivity that are required in medical settings. The outputs from these systems tend to result in serious errors when deployed without human oversight. Such errors are not solely linguistic mistakes but also may pose risks to patient safety and treatment efficacy. This idea reminds us the need for qualified medical translators to maintain their central position in the translation activities with AI being an available toolkit to amplify (Schneiderman, 2020a, as cited in O'Brien, 2023:393) human performance.

Moreover, a translator's technical competence, like being equipped with sufficient skills to use MT technologies effectively and manage translator technologies' instruments like multilingual terminology databases, is important when it comes to ensuring quality outputs in translated texts. Research competence is of the same importance. It enables translators to research and have a grasp on unfamiliar terms, verify pharmaceutical product-related information, and stay up to date with changes in medical guidelines or clinical and pharmaceutical conventions and standards. These skills cannot be deemed optional as they are prerequisites for production of target texts that are clear to the audience, yet, at the same time accurate and compliant.

6.5 Ethical Considerations & Risks Posed by LLMs

With the rise in AI-driven MT workflows there emerge new responsibilities like minding the risks and potential hazards these autonomous systems bear, as emphasized by scholars around the world (cf. Moorkens, 2022; Kimera et al., 2024; Tieber and Baumgarten, 2024; O'Brien, 2024). As these systems progress through various crucial stages, such as data processing, that impact ethical considerations like privacy and consent, -especially for domains such as medicine where individuals' information are highly sensitive-accountability, confidentiality and transparency become major concerns pertaining to all activities that circulate in these stages, including machine translation. That is to say, when we incorporate AI models into high-stakes processes such as healthcare-related processes, there occurs critical ethical considerations concerning patient confidentiality and data

protection (cf. Genovese et al., 2024:13). Therefore, all stakeholders and actuators must be taking mindful and deliberate decisions when structuring a data-driven MT workflow (cf. Kimera et al., 2024:1).

Given the high stakes involved in handling sensitive information and maintaining confidentiality, the implementation of NMT systems carries significant consequences (Kimera et al., 2024:1). This leads to the necessity of examining the ethical framework surrounding AI-based translation technologies, alongside the accountability of their developers (ibid.). AI needs vast amounts of data to be trained on. These data, however, can have consequences and developers are not bounded by law to be accountable if the training data affect patients (cf. Naik et al., 2022:4), thus, there are concerns as to who is responsible in case of such intricacies, because AI has moved beyond the stage where it was nice-to-have, it is rather becoming an integral tool in our technological activities.

Generally, translators have little control over how their translations are reused, especially when their translation memory (TM) or translated texts are used without clear consent in MT system training. Translators may unintentionally contribute to projects or organizations they ethically oppose if their data is reused without transparency or explicit permission (cf. Moorkens, 2022:124-125). Lack of transparency in MT use in professional translation workflows can lead to ethical issues. Translation clients as well as end-users might not be informed adequately about the use of MT, plus, this bring about risks, misunderstandings, and breaches of confidentiality. Ethical practice requires clear communication and agreed-upon standards among all stakeholders.

Since technology advances so quickly and on such a large scale, regulations often lag behind, leaving us dependent on the ethical actions of engineers and developers (cf. Moorkens, 2022:122). But the question is: how, not only technology developers, but all stakeholders can be informed about ethical alignment well enough to eradicate, confront or mitigate potential threats and hazards AI systems pose? This can be achieved through awareness of and adherence to regulatory frameworks and international standards that draw on these concerns and that aim to offer guidance on such concerns.

From, European Union's (EU) Artificial Intelligence Act¹⁸ to Food and Drug Administration's (FDA) nonbinding draft guidance document: "Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products"; from the standard ISO/IEC TR 24368:2022 (Information technology — Artificial intelligence — Overview of ethical and societal concerns)¹⁹ to the "Draft Law"²⁰ proposed to the Turkish Grand National Assembly on 25th June 2024 aiming to regulate AI technology, there are myriad initiatives that have either been proposed, in progress or established across the world both addressing those concerns and also proposing comments and/or imperatives on the use of AI technologies addressing to all parties including developers, consumers, distributors of AI. The ethical concerns to be considered are for everyone involved in translation production process including MT developers, translation buyers, translation agencies, translators and consumers (cf. Moorkens, 2022:137)

There is a view that the versatility and geographical heterogeneity²¹ of standards and frameworks on AI is a broader matter that requires alignment because there is a wide scale of comments these frameworks offer and it can vary from region to region (see: Sankaran, 2025:1). In my view, this is a further scope of AI agenda that needs discussed, hence I will not delve into this side of the discussion, yet, I recommend those interested to consult comparative studies on approaches to AI ethics and AI governance, as they offer valuable insights into the legal, cultural, and ethical aspects shaping the development and deployment of AI technologies across different settings.

To sum up, in the light of the afore-mentioned challenges, a stronger ethical ground is needed for the future development and safe use of AI-driven translation technologies. It is essential that developers, translators, and all other stakeholders included in translation

¹⁸ [https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI\(2021\)698792](https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2021)698792)

¹⁹ <https://www.iso.org/standard/78507.html>

²⁰ https://www.tbmm.gov.tr/Yasama/KanunTeklifi/e21539a0-888a-4500-81be-01904a918c53?TSPD_101_R0=08ffcef486ab2000abc6ac7c8ae52c22502b93e1ff84c1e1306d056722ea9d91f419e6671dbec87708f34eb69c143000a627be6b94a4ff0fb49dc47688dc10bba3d15e87aca9af353840eb715561d9cce4c619d6ba44a689ea64eea8a198cbeb

²¹ By "geographical heterogeneity", it is meant that a number of the guidelines available are region-focused and offer imperatives on a local scale which apply to regional practices. Bodies such as ISO aim to offer international guidelines, nonetheless there is an abundance of region-dependent frameworks currently surrounding "machine ethics".

workflows remain actively engaged with emerging ethical guidelines and regulatory standards so that responsible practices would occur. Increasing awareness on MT-driven translation activities, offering well-designed trainings on AI ethics (both in academia and the industry), and building open communication (among all parties involved in the translated product life-cycle) are crucial steps for minimizing risks and taking accountability. I believe, since global frameworks and regulations have aimed to offer useful guidance, they need to be interpreted and adapted carefully within different sociocultural and legal settings.



7. CONCLUDING REMARKS

This study sought answers to the question of how effective LLMs are in medical translation activities, when it comes to role-based prompting and whether conventional NMT systems, particularly Google Translate, are still effective in medical settings. Based on the findings obtained from all the MT outputs against all three source texts, it would not be wrong to imply that MT cannot be fully relied on when it comes to texts in medical domain, and that role-based prompting does not necessarily enhance quality of translation outputs in medical/pharmaceutical texts domains. It has also been observed that Google Translate does not outperform LLM outputs. Driven by the main question of whether role-based (persona) prompting could notably increase translation quality in medical/pharmaceutical texts, the research has shown both potential advantages and critical limitations in current MT technologies.

The aim was not only to compare performance metrics across different MT outputs, but also to question the values of role of human agency, domain awareness, and ethical responsibility in AI-driven translation activities. The study was conducted with partial references to functionalist translation theories (particularly Nord's and Reiss's propositions) that highlight that translation should be performed taking into account the text's intended function and its target audience. The analysis on text types revealed that while SmPC documents are mainly informative and require technical accuracy for healthcare professionals, PIL documents are blends of informative and operative elements which necessitate readability and accessibility for lay readers.

Despite some models having acceptable BLEU scores across all 3 TT's, both zero-shot outputs and RBP outputs exhibited inconsistencies in terminology, register as well as having made incorrect term usage and unnatural phrasing errors at both minor, major and critical severities. It is worth noting that high BLEU scores did not necessarily mean higher qualitative performance. In some instances, critical translation errors -such as incorrect injection instructions or misrendered clinical concepts- were maintained by the LLMs

regardless of the prompting type. This further demonstrated the significance of human oversight and involvement on qualitative methodology basis.

In medical domain, role-based prompting has -depending on the model, of course- tendency to lead an LLM to make major, even critical errors, as well as zero-shot (raw) prompting. In the pilot experiments carried out within the scope of this study, although RBP outputs partly outperformed their raw outputs, namely had higher BLEU scores at times, when we take error annotation and penalty scores into account, some LLMs have higher penalty scores with their RBP outputs than their raw outputs. It has also been observed that, regardless of the final penalty scores, all MT outputs needed either light or full post-editing, meaning they cannot be directly incorporated into MT workflows unless MTPE has been performed on them. Thus, there hasn't been noted a meaningful impact by role-based prompting in medical domain, at least in the light of this very particular study.

It is worth noting that a translator who is context-aware and has competence in medical/pharmaceutical field may preliminarily perform error annotation on MT outputs, as conducted within the scope of this study, which is a fruitful practice. Subsequently, the translator should analyze first, then compare different MT outputs, not being restrained to one single MT engine or LLM, and determine whether a light or a full-MTPE is necessary in their case. Once the need is determined, the MTPE'd product can be integrated into translation workflow. Translation and MTPE-oriented activities in this domain requires co-existence of linguistic, textual, technical and domain-specific competencies, as stated in standards such as ISO 17100:2015 (cf. Filatova, 2017).

The concept of human-machine collaboration emerged as an alternative to disputes revolving around AI replacing translators. In this thesis, it was aimed to suggest that AI should be regarded not as a replacement but as a partner that humans would benefit in terms of intelligence amplification, which is encouraged to be included in translator's toolkit (cf. Shneiderman, 2020a, as cited in O'Brien, 2024). These tools are meant to amplify human translation activities rather than becoming a substitution.

In addition, the study raised notable ethical questions surrounding the deployment of LLMs in translation workflows, with particular respect to medical domain. Issues such as

data privacy, consent, and re-use of translators' work without explicit consent increasingly emerge. LLMs operate in legally grey areas and regulatory frameworks lag behind technological advancements. This study tried to contribute to the idea that ethical attempts need to be taken for MT workflows by collaboration of translators, developers, institutions and policy-makers, basically all the actors involved in translated product life-cycle.

To sum up, while role-based prompting may offer a structured approach to guide the LLMs in a certain direction, it shares the same limitations as general purpose LLMs in medical domain. Thus, the integration of LLMs into medical translation workflows must be controlled, monitored and needs the involvement of human actuators. The case is the same for conventional NMTs, particularly untrained Google Translate.

Given that current study deploys only pre-trained, ready-to-use AI models, further research on domain-specific (i.e: medical) trained (such as by fine-tuning or by RAG) MT can be conducted to unleash further potentials of role-based prompting. Moreover, in a world where LLMs are rapidly being upgraded to a newer and faster deep-thinking models, with the advent of upcoming models, future research could be conducted to discover whether role-based prompting will make a meaningful impact in medical/pharmaceutical domains thanks to newer models to come.

REFERENCES

- Alimen, Nilüfer. “Makine Çevirisinden Sohbet Robotu Çevirisine: ChatGPT Ile Deneysel Bir Çalışma.” *RumeliDE Dil ve Edebiyat Araştırmaları Dergisi*, no. 36 (October 21, 2023): 1532–48. <https://doi.org/10.29000/rumelide.1369589>.
- Babar, Zia. “The Evolution of Prompt Engineering.” *Medium* (blog), September 30, 2024. <https://medium.com/@zbabar/the-evolution-of-prompt-engineering-3f67d2607473>.
- Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. “Neural Machine Translation by Jointly Learning to Align and Translate.” arXiv, May 19, 2016. <https://doi.org/10.48550/arXiv.1409.0473> arXiv:1409.0473.
- Berghammer, Gabi. “Translation and the language(s) of medicine: Keys to producing a successful German-English translation.” (2006).
- Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. “Language Models Are Few-Shot Learners.” arXiv.org, May 28, 2020. <https://arxiv.org/abs/2005.14165v4>.
- Canbaz, Rüveyda. “Yazılı Tıp Metinleri Çevirisi Sürecinde ChatGPT3 Sohbet Robotunun Konumu ve Çeviri Ürün Kalitesinin Değerlendirilmesi.” Master’s Thesis, Sakarya University, 2024.
- Castilho, Sheila, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. “Is Neural Machine Translation the New State of the Art?” *The Prague Bulletin of Mathematical Linguistics* 108, no. 1 (June 1, 2017): 109–20. <https://doi.org/10.1515/pralin-2017-0013>.
- Çetintaş Sönmez, Esra. “Tıp Metinlerinin Çevirisine İlişkin Yaklaşımlar ve Beklentiler: Bir Olgu Çalışması.” Master’s Thesis, Okan University, 2011.
- Desmond, Michael, and Michelle Brachman. “Exploring Prompt Engineering Practices in the Enterprise.” arXiv, 2024. <https://doi.org/10.48550/ARXIV.2403.08950>.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “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)*, edited by Jill Burstein, Christy Doran, and Tamar Solorio, 4171–86. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. <https://doi.org/10.18653/v1/N19-1423>.

- Ekinci, Sena. "The Effect of Error Annotation on Post-Editing Effort and Post-Edited Product: An Experimental Study on Machine-Translated Subtitles of Educational Content." Master's Thesis, Istanbul 29 Mayıs University, 2022.
- Džuganová, Božena. "English Medical Terminology – Different Ways of Forming Medical Terms." JAHR 4, no. 7 (2013): 55–63. <https://hrcak.srce.hr/file/162511>
- Filatova, Svetlana. "Tıbbi Çevirmen Olmak: Uzmanlar Kimdir? / Becoming a Medical Translator: Who Are the Experts?" Master's Thesis, Istanbul 29 Mayıs University, 2017.
- Fischbach, Henry. "Some Anatomical and Physiological Aspects of Medical Translation: Lexical Equivalence, Ubiquitous References and Universality of Subject Minimize Misunderstanding and Maximize Transfer of Meaning." *Meta: Journal Des Traducteurs / Meta: Translators' Journal* 31, no. 1 (1986): 16–21. <https://doi.org/10.7202/002743ar>.
- Gao, Yuan, Ruili Wang, and Feng Hou. "How to Design Translation Prompts for ChatGPT: An Empirical Study." arXiv, April 21, 2023. <https://doi.org/10.48550/arXiv.2304.02182>. arXiv:2304.02182.
- Genovese, Ariana, Sahar Borna, Cesar A. Gomez-Cabello, Syed Ali Haider, Srinivasagam Prabha, Antonio J. Forte, and Benjamin R. Veenstra. "Artificial Intelligence in Clinical Settings: A Systematic Review of Its Role in Language Translation and Interpretation." *Annals of Translational Medicine* 12, no. 6 (December 2024): 117–117. <https://doi.org/10.21037/atm-24-162>.
- Göktürk, Akşit. "Çeviri: Dillerin Dili." Istanbul: Yapı Kredi Yayınları, 1994.
- Günay Köprülü, Sevtap. "TIBBİ TERMİNOLOJİ VE PROSPEKTÜS (KULLANMA TALİMATI) ÇEVİRİSİ." *Avrasya Terim Dergisi*, 2017.
- Han, Lifeng, Serge Gladkoff, Gleb Erofeev, Irina Sorokina, Betty Galiano, and Goran Nenadic. "Neural Machine Translation of Clinical Text: An Empirical Investigation into Multilingual Pre-Trained Language Models and Transfer-Learning." *Frontiers in Digital Health* 6 (February 26, 2024): 1211564. <https://doi.org/10.3389/fdgth.2024.1211564>.
- International Organization for Standardization and International Electrotechnical Commission. ISO/IEC TR 24368:2022 - Information Technology — Artificial Intelligence — Overview of Ethical and Societal Concerns. Geneva: ISO/IEC, 2022. <https://www.iso.org/standard/78507.html>

- International Organization for Standardization. ISO 17100:2015 – Translation Services – Requirements for Translation Services. Geneva: ISO, 2015. <https://www.iso.org/standard/59149.html>
- Jakobson, Roman. “Closing Statements: Linguistics and Poetics.” In *Style in Language*, edited by Thomas A. Sebeok, 350–377. New York: Wiley, 1960.
- Karwacka, Wioleta. “Quality Assurance in Medical Translation.” *The Journal of Specialised Translation*, no. 21 (2014): 19–34.
- Kimera, Richard, Yun-Seon Kim, and Heeyoul Choi. “Advancing AI with Integrity: Ethical Challenges and Solutions in Neural Machine Translation.” arXiv, April 1, 2024. <https://doi.org/10.48550/arXiv.2404.01070>. arXiv:2404.01070.
- Köprülü, Sevtap Günay. “Tıbbi Terminoloji ve Prospektüs (Kullanma Talimatı) Çevirisi.” *Avrasya Terim Dergisi* 5, no. 1 (2017): 11–18.
- “KULLANMA TALİMATI SKYRIZI 150 Mg/mL Enjeksiyonluk Çözelti İçeren Kullanıma Hazır Kalem.” TİTCK - Türkiye İlaç ve Tıbbi Cihaz Kurumu, June 6, 2024. https://titck.gov.tr/storage/Archive/2024/kubKtAttachments/SKYRIZI150mgmLenjeksiyonlukzeltiierenkullanmahazrkalemkt_41c58123-f33f-4206-b37c-d9f53191073d.pdf.
- “KULLANMA TALİMATI VENCLYXTO 10 Mg Film Kaplı Tablet.” TİTCK - Türkiye İlaç ve Tıbbi Cihaz Kurumu, March 21, 2025. https://titck.gov.tr/storage/Archive/2025/kubKtAttachments/VENCLYXTO10mgfilmkaplitabletkt_24_03_2025.pdf.
- Lyu, Chenyang, Zefeng Du, Jitao Xu, Yitao Duan, Minghao Wu, Teresa Lynn, Alham Fikri Aji, Derek F. Wong, and Longyue Wang. “A Paradigm Shift: The Future of Machine Translation Lies with Large Language Models.” In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, edited by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, 1339–52. Torino, Italia: ELRA and ICCL, 2024. <https://aclanthology.org/2024.lrec-main.120/>.
- Malmkjær, Kirsten, ed. *The Routledge Handbook of Translation Studies and Linguistics*. 1st ed. Routledge, 2017. <https://doi.org/10.4324/9781315692845>.
- Montalt, Vicent, Karen Korning Zethsen, and Wioleta Karwacka. “Medical Translation in the 21st Century – Challenges and Trends.” *MonTi: Monografías de Traducción e Interpretación*, no. 10 (2018): 9–25. <https://doi.org/10.6035/MonTI.2018.10.1>.

- Moorkens, Joss. “Ethics and Machine Translation.” Zenodo, 2022. <https://doi.org/10.5281/ZENODO.6759984>.
- “MQM (Multidimensional Quality Metrics).” Accessed February 7, 2025. <https://themqm.org>.
- Munday, Jeremy. “Introducing Translation Studies: Theories and Applications.” 4th ed. London: Routledge, 2016.
- Naik, Nithesh, B. M. Zeeshan Hameed, Dasharathraj K. Shetty, Dishant Swain, Milap Shah, Rahul Paul, Kaivalya Aggarwal, et al. “Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility?” *Frontiers in Surgery* 9 (March 14, 2022): 862322. <https://doi.org/10.3389/fsurg.2022.862322>.
- O’Brien, Sharon. “Human-Centered Augmented Translation: Against Antagonistic Dualisms.” *Perspectives* 32, no. 3 (May 3, 2024): 391–406. <https://doi.org/10.1080/0907676X.2023.2247423>.
- Oğuz, Derya. “Translation Problems in Medical Texts.” *The Journal of International Social Research* 7, no. 34 (2014): 127–134.
- “OPDIVO 10 Mg/mL Concentrate for Solution for Infusion - Summary of Product Characteristics (SmPC) - (Emc) | 6888.” Accessed June 4, 2025. <https://www.medicines.org.uk/emc/product/6888/smpc>.
- “OPDIVO 40 Mg/4 mL IV İnfüzyonluk Çözelti Konsantresi İçeren Flakon – Kısa Ürün Bilgisi.” TİTCK - Türkiye İlaç ve Tıbbi Cihaz Kurumu, 2022. https://www.titck.gov.tr/storage/Archive/2022/kubKtAttachments/TTCKOnaylKTOPDIVO40MG_360f4db8-28f9-414e-8197-4bbb031793a7.pdf.
- Öner Bulut, Senem. “Integrating Machine Translation into Translator Training: Towards ‘Human Translator Competence’?” *transLogos Translation Studies Journal* 2/2, no. 2/2 (2019): 1–26. <https://doi.org/10.29228/transLogos.11>.
- Öner, Işın, and Zehra Begüm Bengi. “Essential or Obsolete? The Role of Human Competencies in the Tech-Driven Language Services Industry.” *transLogos Translation Studies Journal* 7/1, no. 7/1 (2024): 78–104. <https://doi.org/10.29228/transLogos.66>.
- Öner, Işın, and Senem Öner Bulut. “Post-Editing Oriented Human Quality Evaluation of Neural Machine Translation in Translator Training: A Study on Perceived Difficulties and Benefits.” *transLogos Translation Studies Journal* 4, no. 1 (2021): 100–124. <https://doi.org/10.29228/transLogos.33>.

- Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "BLEU: A Method for Automatic Evaluation of Machine Translation." In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 311. Philadelphia, Pennsylvania: Association for Computational Linguistics, 2001. <https://doi.org/10.3115/1073083.1073135>.
- Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language Models Are Unsupervised Multitask Learners." *OpenAI Blog*, February 2019. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. "COMET: A Neural Framework for MT Evaluation." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2685–2702. Online: Association for Computational Linguistics, 2020. <https://doi.org/10.18653/v1/2020.emnlp-main.213>.
- Reiß, Katharina, and Hans J. Vermeer. "Towards a General Theory of Translational Action: Skopos Theory Explained." Translated by Christiane Nord. English reviewed by Marina Dudenhöfer. London: Routledge, 2014.
- Reynolds, Laria, and Kyle McDonell. "Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm." arXiv, February 15, 2021. <https://doi.org/10.48550/arXiv.2102.07350>. arXiv:2102.07350.
- Sahoo, Pranab, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. "A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications." arXiv, 2024. <https://doi.org/10.48550/ARXIV.2402.07927>.
- Sankaran, Sridharan. "Enhancing Trust Through Standards: A Comparative Risk-Impact Framework for Aligning ISO AI Standards with Global Ethical and Regulatory Contexts." arXiv, April 22, 2025. <https://doi.org/10.48550/arXiv.2504.16139>. arXiv:2504.16139.
- Schumacher, Perrine. "Fostering MT Literacy and Reasserting the Value of Human Translators." *transLogos Translation Studies Journal* 6/2, no. 6/2 (2023): 1–20. <https://doi.org/10.29228/transLogos.57>.
- Scius-Bertrand, Anna, Michael Jungo, Lars Vögtlin, Jean-Marc Spat, and Andreas Fischer. "Zero-Shot Prompting and Few-Shot Fine-Tuning: Revisiting Document Image Classification Using Large Language Models." arXiv, December 18, 2024. <https://doi.org/10.48550/arXiv.2412.13859>. arXiv:2412.13859.

- Seçkin, Sevcan, and Emine Bogenç Demirel. “Kurumsal Alanın Gerektirdiği Çevirmen Modeli.” *Ege Sosyal Bilimler Dergisi* 2, no. 2 (2019): 109–122.
- Shanahan, Murray, Kyle McDonell, and Laria Reynolds. “Role Play with Large Language Models.” *Nature* 623, no. 7987 (November 16, 2023): 493–98. <https://doi.org/10.1038/s41586-023-06647-8>.
- Sharkey, Edward, and Philip Treleaven. “BERT vs GPT for Financial Engineering.” arXiv, 2024. <https://doi.org/10.48550/ARXIV.2405.12990>.
- “Skyrizi 150 Mg Solution for Injection in Pre-Filled Pen.” EMC - electronic medicines compendium, n.d. <https://www.medicines.org.uk/emc/product/12625/pil#gref>.
- Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. “Sequence to Sequence Learning with Neural Networks.” *Advances in Neural Information Processing Systems* 27 (2014): 3104–3112.
- Taşdan Doğan, Tuğçe Elif. “Bourdieu Sosyolojisi Işığında Tıbbi Metin Çevirileri: Covid-19 Örneği.” *RumeliDE Dil ve Edebiyat Araştırmaları Dergisi*, no. 21 (2020): 933–948.
- Tieber, Michael, and Stefan Baumgarten. “Mean Machines? Sociotechnical (r)Evolution and Human Labour in the Translation and Interpreting Industry.” *Perspectives* 32, no. 3 (May 3, 2024): 379–90. <https://doi.org/10.1080/0907676X.2024.2346378>.
- Tracey, David. “An Early History of Medical Translation.” *JDDG: Journal Der Deutschen Dermatologischen Gesellschaft* 16, no. 10 (October 2018): 1300–1301. <https://doi.org/10.1111/ddg.13667>.
- Turkish Law Blog. “Review of Turkey’s Draft Artificial Intelligence Law.” Turkish Law Blog, July 8, 2024. <https://turkishlawblog.com/insights/detail/review-of-turkeys-draft-artificial-intelligence-law>
- Türkiye Büyük Millet Meclisi. Yapay Zekâ Alanında Bazı Düzenlemeler Yapılmasına Dair Kanun Teklifi. Ankara: Türkiye Büyük Millet Meclisi, 2024.
- U.S. Food and Drug Administration. “Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products: Draft Guidance for Industry.” Silver Spring, MD: FDA, January 2025. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological>.

- Unbabel. “Unbabel/wmt22-comet-da.” Hugging Face, 2022.
<https://huggingface.co/Unbabel/wmt22-comet-da>
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. “Attention Is All You Need.” arXiv, August 2, 2023. <https://doi.org/10.48550/arXiv.1706.03762>. arXiv:1706.03762.
- “Venclyxto 100 Mg Film-Coated Tablets.” EMC - electronic medicines compendium, n.d.
<https://www.medicines.org.uk/emc/product/10476/pil#gref>.
- Vöge, Lukas, Vincent Gurgul, and Stefan Lessmann. “Leveraging Zero-Shot Prompting for Efficient Language Model Distillation.” arXiv, March 23, 2024.
<https://doi.org/10.48550/arXiv.2403.15886>. arXiv:2403.15886.
- Yamada, Masaru. “Optimizing Machine Translation through Prompt Engineering: An Investigation into ChatGPT’s Customizability.” arXiv, February 21, 2024.
<https://doi.org/10.48550/arXiv.2308.01391>. arXiv:2308.01391.
- Yagbasan, Mustafa, and İhsan Kurtbaş. “Bir Yazılı Anlatım Türü Olan İlaç Prospektüslerindeki Dil ve Anlam Sorununa Alan Araştırması Yöntemiyle Eleştirel Bir Bakış.” *Kocaeli Üniversitesi İletişim Fakültesi Araştırma Dergisi*, no. 7 (2006): 100–121.
- Zhang, Biao, Barry Haddow, and Alexandra Birch. “Prompting Large Language Model for Machine Translation: A Case Study.” arXiv.org, January 17, 2023.
<https://arxiv.org/abs/2301.07069v2>.
- Zheng, Mingqian, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. “When ‘A Helpful Assistant’ Is Not Really Helpful: Personas in System Prompts Do Not Improve Performances of Large Language Models.” arXiv, October 9, 2024.
<https://doi.org/10.48550/arXiv.2311.10054>. arXiv:2311.10054.

APPENDIX A – EXTRACT FROM ST-1 (PAGE 12)

5.1 Pharmacodynamic properties

Pharmacotherapeutic group: Antineoplastic agents, monoclonal antibodies. ATC code: L01XC17.

Mechanism of action

Nivolumab is a human immunoglobulin G4 (IgG4) monoclonal antibody (HuMAb), which binds to the programmed death-1 (PD-1) receptor and blocks its interaction with PD-L1 and PD-L2. The PD-1 receptor is a negative regulator of T-cell activity that has been shown to be involved in the control of T-cell immune responses. Engagement of PD-1 with the ligands PD-L1 and PD-L2, which are expressed in antigen presenting cells and may be expressed by tumours or other cells in the tumour microenvironment, results in inhibition of T-cell proliferation and cytokine secretion. Nivolumab potentiates T-cell responses, including anti-tumour responses, through blockade of PD-1 binding to PD-L1 and PD-L2 ligands. In syngeneic mouse models, blocking PD-1 activity resulted in decreased tumour growth.

Clinical efficacy and safety

Randomised phase 3 study vs. dacarbazine (CA209066)

The safety and efficacy of nivolumab 3 mg/kg for the treatment of advanced (unresectable or metastatic) melanoma were evaluated in a phase 3, randomised, double-blind study (CA209066). The study included adult patients (18 years or older) with confirmed, treatment-naïve, Stage III or IV BRAF wild-type melanoma and an Eastern Cooperative Oncology Group (ECOG) performance-status score of 0 or 1. Patients with active autoimmune disease, ocular melanoma, or active brain or leptomeningeal metastases were excluded from the study.

A total of 418 patients were randomised to receive either nivolumab (n = 210) administered intravenously over 60 minutes at 3 mg/kg every 2 weeks or dacarbazine (n = 208) at 1000 mg/m² every 3 weeks. Randomisation was stratified by PD-L1 status and M stage (M0/M1a/M1b versus M1c). Treatment was continued as long as clinical benefit was observed or until treatment was no longer tolerated. Treatment after disease progression was permitted for patients who had a clinical benefit and did not have substantial adverse effects with the study drug, as determined by the investigator. Tumour assessments, according to the Response Evaluation Criteria in Solid Tumours (RECIST), version 1.1, were conducted 9 weeks after randomisation and continued every 6 weeks for the first year and then every 12 weeks thereafter. The primary efficacy outcome measure was overall survival (OS). Key secondary efficacy outcome measures were investigator-assessed progression-free survival (PFS) and objective response rate (ORR).

Baseline characteristics were balanced between the two groups. The median age was 65 years (range: 18-87), 59% were men, and 99.5% were white. Most patients had ECOG performance score of 0 (64%) or 1 (34%). Sixty-one percent of patients had M1c stage disease at study entry. Seventy-four percent of patients had cutaneous melanoma, and 11% had mucosal melanoma; 35% of patients had PD-L1 positive melanoma (≥5% tumour cell membrane expression). Sixteen percent of patients had received prior adjuvant therapy; the most common adjuvant treatment was interferon (9%). Four percent of patients had a history of brain metastasis, and 37% of patients had a baseline LDH level greater than ULN at study entry.

The Kaplan-Meier curves for OS are shown in Figure 1.

APPENDIX B – EXTRACT FROM REFERENCE TEXT-1 (PAGE 30)

Klinik etkililik ve güvenlik

Doz/maruziyet etkililik ve güvenlik ilişkilerinin modellenmesine dayanarak, her 2 haftada bir 240 mg veya her 2 haftada bir 3 mg/kg nivolumab arasında etkililik ve güvenlik açısından klinik olarak anlamlı bir fark bulunmamaktadır. Ek olarak, bu ilişkilere dayanarak, RHK'de 4 haftada bir 480 mg veya her 2 haftada bir 3 mg/kg nivolumab dozu arasında klinik olarak anlamlı bir fark görülmemiştir.

Melanom

İlerlemiş melanom tedavisi

Dakarbazin karşısında randomize faz 3 çalışma (CA209066)

Nivolumab 3 mg/kg'ın ilerlemiş (rezeke edilemeyen veya metastatik) melanom tedavisi için güvenliliği ve etkililiği bir faz 3, randomize, çift kör çalışmada (CA209066) değerlendirilmiştir. Çalışmaya doğrulanmış, tedavi uygulanmamış, Evre III veya IV BRAF vahşi tip melanomu olan ve ECOG performans durumu 0 veya 1 olan yetişkin hastalar (18 yaş ve üzeri) dahil edilmiştir. Aktif otoimmün hastalığı, oküler melanomu veya aktif beyin ya da leptomeningeal metastazları bulunan hastalar çalışmaya alınmamıştır.

Toplam 418 hasta her 2 haftada bir 60 dakika süreyle intravenöz yoldan uygulanan 3 mg/kg dozunda nivolumab (n = 210) veya her 3 haftada bir 1000 mg/m² dozunda dakarbazin (n = 208) almak üzere randomize edilmiştir. Randomizasyon, tümör PD-L1 durumu ve M evresi (M1c karşısında M0/M1a/M1b) ile basamaklandırılmıştır. Tedavi, klinik yarar gözlemlendiği sürece veya tedavi hasta tarafından artık tolere edilemeyinceye kadar sürdürülmüştür. Hastalık progresyonundan sonra tedaviye, araştırmacı tarafından belirlendiği gibi klinik yarar gören ve çalışma ilacı ile önemli bir advers olay yaşamamış olan hastalarda izin verilmiştir. Tümör değerlendirmeleri, randomizasyondan 9 hafta sonra, ardından ilk yıl her 6 haftada bir ve bunu takip eden yıllarda her 12 haftada bir, Solid Tümörlerde Yanıt Değerlendirme Kriterleri (RECIST), versiyon 1.1 uyarınca gerçekleştirilmiştir. Primer etkililik sonlanım noktası genel sağkalımdı (OS). Kilit sekonder etkililik sonlanım noktası ölçümleri, araştırmacı tarafından değerlendirilen progresyonsuz sağkalım (PFS) ve objektif yanıt oranı (ORR) idi.

İki grup arasında başlangıç özellikleri dengelenmişti. Yaş ortalaması 65 olup (aralık: 18 - 87) %59'u erkek ve %99,5'i beyazdı. Başlangıç ECOG performans durumu puanı 0 (%64) veya 1 (%34) idi. Hastaların %61'i, çalışmaya girişte evre M1c hastalığa sahipti. %74'ü kutanöz melanom, %11'i mukozal melanom; %35'i ise PD-L1 pozitif melanoma sahipti ($\geq$ %5 tümör hücresi membran ekspresyonu). Hastaların %16'sı, daha önce adjuvan tedavi görmüş olup en yaygın kullanılan adjuvan tedavi interferondur (%9). Hastaların %4'ünde çalışma girişinde beyin metastazı, %37'sinde ise ULN'den yüksek LDH düzeyi mevcuttu.

OS için Kaplan-Meier eğrileri Şekil 1'de gösterilmektedir.

APPENDIX C – EXTRACT FROM ST-2 (PAGE 4)

- Tell your doctor immediately if you become pregnant while you are taking this medicine.

Breast-feeding

Do not breast-feed while you are taking this medicine. It is not known whether the active substance in Venclyxto passes into breast milk.

Fertility

Based on findings in animals, Venclyxto may cause male infertility (low or no sperm count). This may affect your ability to father a child. Ask your doctor for advice on sperm storage before starting treatment with Venclyxto.

Driving and using machines

You may feel tired or dizzy after taking Venclyxto, which may affect your ability to drive or use tools or machines. If this happens, do not drive or use any tools or machines.

Venclyxto contains sodium

This medicine contains less than 1 mmol sodium (23 mg) per tablet, that is to say essentially “sodium free”.

3. How to take Venclyxto

Always take this medicine exactly as your doctor, pharmacist, or nurse has told you. Check with your doctor, pharmacist, or nurse if you are not sure.

How much to take

If you have CLL

You will begin treatment with Venclyxto at a low dose for 1 week. Your doctor will gradually increase the dose over the next 4 weeks to the full standard dose. For the first 4 weeks you will get a new pack each week.

- the starting dose is 20 mg (two 10 mg tablets) once a day for 7 days.
- the dose will be increased to 50 mg (one 50 mg tablet) once a day for 7 days.
- the dose will be increased to 100 mg (one 100 mg tablet) once a day for 7 days.
- the dose will be increased to 200 mg (two 100 mg tablets) once a day for 7 days.
- the dose will be increased to 400 mg (four 100 mg tablets) once a day for 7 days.
 - When you are receiving Venclyxto therapy alone, you will stay on the 400 mg daily dose, which is the standard dose, for as long as necessary.
 - When you are receiving Venclyxto therapy in combination with rituximab, you will receive the 400 mg daily dose for 24 months.
 - When you are receiving Venclyxto therapy in combination with obinutuzumab, you will receive the 400 mg daily dose for approximately 10 months.
 - When you are receiving Venclyxto therapy in combination with ibrutinib, you will receive the 400 mg daily dose for approximately 10 months.

Your dose may need to be adjusted for side effects. Your doctor will advise what your dose should be.

If you have AML

You will begin treatment with Venclyxto on a lower dose. Your doctor will gradually increase the dose each day for the first 3 or 4 days (depending on what medicine you take Venclyxto with). After 3 or 4 days you will take the full standard dose. The dose (tablets) is taken once a day.

APPENDIX D – EXTRACT FROM REFERENCE TEXT-2 (PAGE 4)

AML hastasıysanız;

AML hastasıysanız hastanede tedavi görmeniz gerekebilir ve doktorunuz veya hemşireniz yeterli su/sıvı almanızı sağlayacaktır, size vücudunuzda ürik asit birikmesini engelleyen ilaçlar verecektir ve siz VENCLYXTO almaya başlamadan önce, dozunuzu artırırken ve tam dozu almaya başladığınızda kan testleri yapacaktır.

Bu uyarılar, geçmişteki herhangi bir dönemde dahi olsa sizin için geçerliyse lütfen doktorunuza danışınız.

VENCLYXTO’nun yiyecek ve içecek ile kullanılması

VENCLYXTO kullanırken greyfurt ürünleri, turuncgiller (Sevilya portakalı, acı portakal) ya da yıldız meyvesi (karambola) yemeyiniz- buna bu meyvelerin yenilmesi, meyve sularının içilmesi veya bunları içeren destekleyicilerin alınması da dahildir. Bu ürünler VENCLYXTO’nun kanınızdaki miktarını artırabilir.

Hamilelik

İlacı kullanmadan önce doktorunuza veya eczacınıza danışınız.

- Bu ilacı kullanırken gebe kalmayınız. Eğer gebe iseniz, gebe olabileceğinizi düşünüyorsanız veya gebe kalmayı planlıyorsanız bu ilacı almadan önce doktorunuza, eczacınıza ya da hemşirenize danışınız.
- VENCLYXTO gebelik süresince kullanılmamalıdır. VENCLYXTO’nun gebe kadınlarda güvenliliğine ait bilgiler bulunmamaktadır.
- Çocuk doğurma potansiyeline sahip kadınların, gebelikten korunmak için tedavi süresince ve VENCLYXTO kullandıktan sonra en az 30 gün boyunca yüksek etkili bir doğum kontrol yöntemi kullanması gerekir. Eğer hormonal doğum kontrol hapları veya araçları kullanıyorsanız, VENCLYXTO hormonal doğum kontrol haplarının veya araçlarının etkililiğine etki edebileceğinden, ayrıca bariyer doğum kontrol metodu da kullanmalısınız (kondom gibi).
- Bu ilacı kullanırken gebe kalırsanız bu durumu doktorunuza hemen bildiriniz.

Tedaviniz sırasında hamile olduğunuzu fark ederseniz hemen doktorunuza veya eczacınıza danışınız.

Emzirme

İlacı kullanmadan önce doktorunuza veya eczacınıza danışınız.

Bu ilacı kullanırken emzirmeyiniz. VENCLYXTO’nun etkin maddesinin anne sütüne geçip geçmediği bilinmemektedir. Emzirilen çocuk açısından bir risk olduğu göz ardı edilemez. VENCLYXTO ile tedavi sırasında emzirme durdurulmalıdır.

Fertilite (üreme yeteneği)

Hayvan çalışmaları bulgularına göre, VENCLYXTO erkek infertilitesine (sperm sayısında düşme ya da sperm yokluğu) neden olabilir. Bu durum baba olma yetinizi etkileyebilir. VENCLYXTO tedavisine başlamadan önce sperm depolama hakkında doktorunuza danışınız.

Araç ve makine kullanımı

VENCLYXTO kullandıktan sonra, araç ve makine kullanımı kabiliyetinizi etkileyebilecek yorgunluk veya baş dönmesi hissedebilirsiniz.

APPENDIX E – EXTRACT FROM ST-3 (PAGES 6-7)

- You should receive training on how to inject Skyrizi before giving an injection. Talk to your doctor, pharmacist or nurse if you need help
- Mark the dates on your calendar so you know when to inject Skyrizi
- Keep Skyrizi in the original carton to protect the medicine from light until it is time to use it
- Take the carton out of the refrigerator and leave it at room temperature, out of direct sunlight, for **30 to 90 minutes** before injecting
- **Do not** inject if the liquid in the inspection window is cloudy or contains flakes or large particles. The liquid should look clear to yellow and may contain tiny white or clear particles
- **Do not** shake the pen
- Wait to remove the dark grey cap until just before the injection

Return this medicine to the pharmacy

- if the expiry date (EXP) has passed
- if the liquid has ever been frozen (even if thawed)
- if the pen has been dropped or damaged
- if the carton perforations are broken

Follow the steps below each time you use Skyrizi

STEP 1



Take the carton out of the refrigerator and leave it at room temperature, out of direct sunlight, for **30 to 90 minutes** before injecting.

- **Do not** remove the pen from the carton while allowing Skyrizi to reach room temperature
- **Do not** warm Skyrizi in any other way. For example, **do not** warm it in a microwave or in hot water
- **Do not** use the pen if liquid has been frozen, even if it has been thawed

STEP 2



Place the following items on a clean, flat surface:

- 1 pre-filled pen
- 1 alcohol pad (not included in the carton)
- 1 cotton ball or gauze pad (not included in the carton)
- special disposal container (not included in the carton)

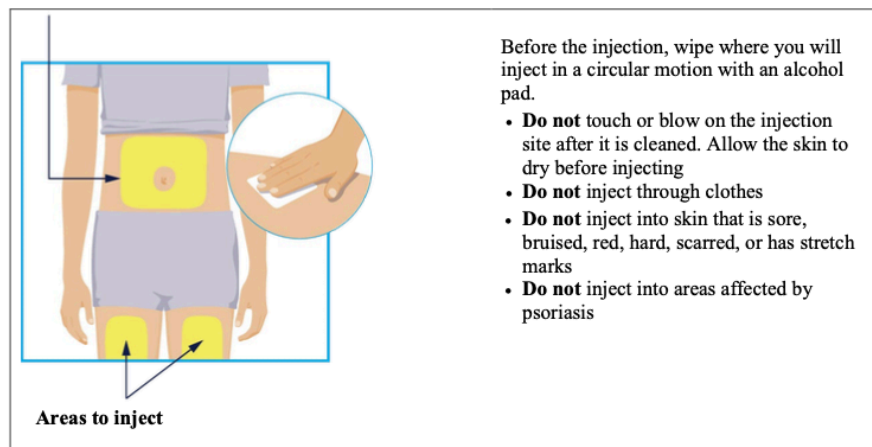
Wash and dry your hands.

STEP 3

Areas to inject

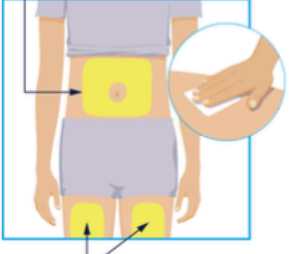
Choose from these 3 areas to inject:

- front of left thigh
- front of right thigh
- your belly (abdomen) at least 5 cm from your belly button (navel)



APPENDIX F – EXTRACT FROM REFERENCE TEXT-3 (PAGE 6)

SKYRIZI'yi her kullanışınızda aşağıdaki adımları takip ediniz

ADIM 1 	Kutuyu buzdolabından çıkarınız ve enjeksiyondan önce 30 ila 90 dakika boyunca, doğrudan güneş ışığı almayacak şekilde oda sıcaklığında bekletin. • Skyrizi'nin oda sıcaklığına gelmesini beklerken kalemi kutusundan çıkarmayınız. • Skyrizi başka bir şekilde ısıtmayınız. Örneğin; bir mikrodalga ya da sıcak su ile ısıtmayınız. • Sıvı donmuşsa, eritilse bile, kalemi kullanmayınız.
ADIM 2 	Aşağıdaki malzemeleri temiz ve düz bir yüzey üzerine yerleştiriniz: • 1 adet kullanıma hazır kalem • 1 adet alkollü ped (kutuya dahil değildir) • 1 adet pamuk yumağı veya gazlı bez (kutuya dahil değildir) • Özel atık kabı (kutuya dahil değildir) Ellerinizi yıkayıp kurulayınız.
ADIM 3 Enjeksiyon yapılacak bölgeler 	Enjeksiyon için şu 3 bölgeden birini seçiniz: • sol uyluğun önü • sağ uyluğun önü • göbek deliğinin en az 5 cm uzağında olmak üzere göbek bölgesi (karın) Enjeksiyondan önce, enjeksiyon yapılacak yeri bir alkollü ped ile dairesel hareketlerle siliniz. • Temizledikten sonra enjeksiyon bölgesine dokunmayınız ya da üfleme yapmayınız. Enjeksiyondan önce derinizin kurumasını bekleyiniz. • Kıyafetlerin üstünden enjeksiyon yapmayınız. • Ağrıyan, morarmış, kızarmış, sertleşmiş, yaralı veya çatlak izleri olan deri üzerine enjeksiyon yapmayınız. • Sedef hastalığından etkilenmiş bölgelere enjeksiyon yapmayınız.

Belge D... Bu belge, güvenli elektronik imza ile imzalanmıştır. W56ZmxXS3k0YnUy Belge Tarih Adresi: https://www.turkiye.gov.tr/saglik-titck-ebys

ÖZGEÇMİŞ

Adı, Soyadı	Mehmet Ferit		DOĞAN
Doğum Yeri ve Yılı			
Bildiği Yabancı Diller	İngilizce		Fransızca
ve Düzeyi	İleri		Başlangıç
Eğitim Durumu	Başlama - Bitirme Yılı	Kurum Adı	
Lise	2012	2016	Alanya Nimet Alaettinoğlu Anadolu Lisesi
Lisans	2016	2021	Marmara Üniversitesi
Yüksek Lisans	2022	2025	İstanbul 29 Mayıs Üniversitesi
Doktora			
Çalıştığı Kurum/lar	Başlama - Ayrılma Yılı	Çalışılan Kurumun Adı	
1.	2022	2024	Okuyan Gümrük Müşavirliği
2.	2024	2024	Quadro CRO Klinik Araştırma
3.	2024	Devam Ediyor	Beigene Turkey Tıbbi Ürünler
Üye Olduğu Bilimsel ve Mesleki Kuruluşlar			
Katıldığı Proje ve Toplantılar			
Yayınlar:			
Diğer:			
İletişim (e-posta):			
Tarih		03.06.2025	
İmza			
Adı Soyadı		Mehmet Ferit DOĞAN	

