Qədim Diyar Avqust 2026, Cild: 8, Sayı: 8, s. 16–33
Ancient Land August 2026, Volume: 8, Issue: 8, pp. 16–33
DOI: https://doi.org/10.36719/2706-6185/62/16-33

Hallucination in Turkish–Arabic LLM Translation Outputs: A Text-Type Based Translation Studies Analysis

Sezer Yılmaz

Abstract. *Hallucination, which has emerged as a significant research problem in machine translation literature, refers to a phenomenon in which large language models (LLMs), despite producing fluent and seemingly coherent translation outputs, may distort source-text information, generate information not found in the source, or produce outputs with a weak semantic connection to it. This study examines hallucination types that may arise in turkish – arabic machine translation from a translation studies perspective. Considering the morphological and syntactic characteristics of Turkish and Arabic, this language pair offers a productive and distinctive field of inquiry for discussing hallucination in translation. The study analyzes three sample texts representing literary, legal, and medical text types through outputs generated by ChatGPT, Grok, and Gemini. The translation outputs were qualitatively examined in terms of fidelity to the source text, semantic integrity, terminological accuracy, source-external additions, task-irrelevant explanations, and repetition. The qualitative findings were interpreted in relation to BERTScore results calculated in Google Colaboratory. BERTScore was treated as a complementary indicator of semantic proximity between the machine translation outputs and the human reference translation, while the final identification and classification of hallucination and deviation types were based on qualitative comparison with the source text. The analyses show that a high semantic similarity score does not always mean that the translation is free from hallucination. In some outputs, even when the score is high, task-irrelevant explanations, interpretive expansions, terminological deviations, or repetitions can still be found. For this reason, automatic metrics should not be used alone when evaluating hallucinated translation outputs. In this study, hallucination is examined together with source-text fidelity, automatic evaluation metrics, and text type.*

Keywords: *hallucination, machine translation, large language models, turkish–arabic translation, BERTScore, translation studies*

Kırıkkale University, PhD in Philology, Kırıkkale, Türkiye
E-mail: sezeryilmaz@kku.edu.tr
Received: 7 April 2026; Accepted: 10 July 2026; Published online: 28 August 2026

Türk–ərəb BDM tərcümə çıxışlarında halüsinasiya: mətn növü əsaslı tərcüməşünaslıq təhlili

Sezer Yılmaz

Xülasə. *Halüsinasiya maşın tərcüməsi ədəbiyyatında mühüm tədqiqat problemi kimi önə çıxır. Bu anlayış böyük dil modellərinin (BDM) axıcı və zahirən tutarlı tərcümə çıxışları yaratmasına baxmayaraq, mənbə mətndəki informasiyanı təhrif etməsi, mənbədə olmayan informasiya əlavə etməsi və ya mənbə mətnlə zəif semantik əlaqəyə malik çıxışlar yaratması ilə bağlıdır. Bu tədqiqat türk–ərəb maşın tərcüməsində ortaya çıxa biləcək halüsinasiya növlərini tərcüməşünaslıq baxımından araşdırır. Türk və ərəb dillərinin morfoloji və sintaktik xüsusiyyətləri nəzərə alındıqda, bu dil cütü tərcüməsində halüsinasiya məsələsinin müzakirəsi üçün məhsuldar və fərqli bir tədqiqat sahəsi təqdim edir. Tədqiqatda ədəbi, hüquqi və tibbi mətn növlərini təmsil edən üç nümunə mətn ChatGPT, Grok və Gemini tərəfindən yaradılan çıxışlar əsasında təhlil edilmişdir. Tərcümə çıxışları mənbə mətnə sadiqlik, semantik bütövlük, terminoloji dəqiqlik, mənbədənkənar əlavələr, tapşırıqdan kənar izahlar və təkrarlar baxımından keyfiyyətcə araşdırılmışdır. Keyfiyyət təhlilinin nəticələri Google Colaboratory mühitində hesablanan BERTScore nəticələri ilə birlikdə şərh edilmişdir. BERTScore maşın tərcüməsi çıxışları ilə insan tərəfindən hazırlanmış istinad tərcüməsi arasındakı semantik yaxınlığı göstərən tamamlayıcı göstərici kimi istifadə edilmiş, halüsinasiya və sapma növlərinin yekun müəyyənləşdirilməsi və təsnifatı isə mənbə mətnlə aparılan keyfiyyət müqayisəsinə əsaslanmışdır. Analizlər göstərir ki, yüksək semantik oxşarlıq göstəricisi tərcümənin halüsinasiyadan tamamilə uzaq olduğu anlamına gəlmir. Bəzi çıxışlarda göstərici yüksək olsa belə, tapşırıqdan kənar izahlar, şərhedici genişləndirmələr, terminoloji sapmalar və ya təkrarlar müşahidə oluna bilər. Bu səbəbdən avtomatik metriklər halüsinativ tərcümə çıxışlarını qiymətləndirərkən təkbaşına istifadə edilməməlidir. Bu tədqiqatda halüsinasiya mənbə mətnə sadiqlik, avtomatik qiymətləndirmə metrikləri və mətn növü ilə birlikdə araşdırılır.*

Açar sözlər: *halüsinasiya, maşın tərcüməsi, böyük dil modelləri, türk–ərəb tərcüməsi, BERTScore, tərcüməşünaslıq*

Kırıkkale Universiteti, filologiya üzrə fəlsəfə doktoru, Kırıkkale, Türkiyə
E-poçt: sezeryilmaz@kku.edu.tr
Daxil oldu: 7 Aprel 2026; Qəbul edildi: 10 İyul 2026; Onlayn dərc edildi: 28 Avqust 2026

Introduction

Translation has long been one of the main ways of enabling communication between different languages, societies, and cultures. Since translation involves carrying meaning from one language and cultural context into another, its quality is often discussed in translation studies through the concept of equivalence. In this sense, the relationship between the source text and the target text is evaluated not only through semantic correspondence, but also through context and function. Whether machine-generated translations can achieve this kind of equivalence is still an important question.

The development of large language models (LLMs) has changed the field of natural language processing (NLP), especially in areas such as language understanding, text generation, reasoning, and access to information (Huang et al., 2025). One area directly affected by this development is natural language generation. Natural language generation includes tasks such as summarization,

dialogue generation, generative question answering, data-to-text generation, and machine translation. For this reason, it is considered an important but challenging subfield of NLP (Ji et al., 2023).

However, the texts produced by these systems do not always meet expectations in terms of semantic accuracy, contextual appropriateness, and consistency with the user's input. Although large language models can perform well in many tasks, they may sometimes go beyond the given prompt and produce content that is inconsistent with the previous context or factual knowledge. This problem is generally referred to in the literature as hallucination (Zhang et al., 2025).

Hallucination is one of the major issues that limits the reliability of large language models. These models can produce outputs that appear fluent, coherent, and convincing, even when the information they contain is not factually grounded. This makes hallucinated content difficult to detect, especially for users who do not have sufficient knowledge of the subject matter. Since such outputs may influence decision-making processes, they may also lead to misinformation and raise ethical concerns (Huang et al., 2025).
The increasing use of artificial intelligence and natural language processing technologies has also made automatic translation systems more accessible. The shift from neural machine translation systems to large language models has made translation faster and easier to use. However, this development has also brought new questions concerning translation quality, equivalence, and reliability. From this perspective, hallucination should not be considered only as a technical problem, but also as an issue that needs to be discussed within translation studies.

Studies on hallucination in machine translation generally focus on widely studied language pairs and translation directions. This creates a limited view of how hallucination may appear in language pairs with different morphological and syntactic structures. Turkish and Arabic, as morphologically rich languages, require particular attention in this regard. Therefore, examining hallucination in Turkish–Arabic machine translation can contribute to a better understanding of how hallucinated deviations emerge in this language pair.

Accordingly, this study examines the types of hallucination that may arise in Turkish–Arabic machine translation from a translation studies perspective. Focusing on literary, legal, and medical texts, it discusses hallucination in relation to fidelity to the source text, semantic and contextual adequacy, terminological accuracy, target-language norms, and the limitations of automatic evaluation approaches.

1. Theoretical Framework

1.1. Equivalence in Translation Studies

To explain why hallucination is a significant issue in translation, it is useful to begin with the concept of equivalence, which is one of the central concepts of translation studies. From this perspective, translation quality is not limited to grammatical correctness. It is also related to how far meaning, context, and function are preserved between the source and target texts. For this reason, the relationship between the source text and the target text is often discussed through the concept of equivalence.

Equivalence refers to the transfer of the meaning in the source text into the target language through different linguistic and structural means (Gürçağlar, 2016). This concept has been explained in different ways in translation studies. Nida (2000) distinguishes between formal equivalence and dynamic equivalence, while also emphasizing that absolute correspondence between languages is not possible. Toury (1980), on the other hand, relates equivalence to the norms governing the translation process. Baker (2018) discusses equivalence at different levels, including word level, above-word

level, grammatical, textual and pragmatic levels. She also states that equivalence is relative and can be affected by linguistic and cultural factors.

Although equivalence has been approached from different theoretical perspectives, the question of whether automatic translation systems can achieve it reliably remains important. This is because machine-generated outputs may sometimes include elements that are not present in the source text. Such outputs may appear fluent and acceptable at first sight, but they can still create serious problems in terms of semantic equivalence. Therefore, hallucination becomes relevant to translation studies at the point where surface fluency conflicts with fidelity to the source text. In this respect, hallucination can be considered a violation of equivalence.

1.2. Quality and Ethics in Machine Translation

Machine translation is a translation process carried out through artificial intelligence-based algorithms and systems based on these algorithms in order to transfer meaning between languages. The widespread use of artificial intelligence systems increases the importance of ethical principles such as transparency, accountability, cultural sensitivity, and human oversight (Kimera et al., 2024). From a Translation Studies perspective, translation ethics is addressed primarily in relation to the translator and encompasses the moral and professional dimensions of the decisions made throughout the translation process. In this respect, translation ethics is closely associated with the translator's professional autonomy, responsibility, and professional conduct. However, with the integration of artificial intelligence into translation processes, ethical considerations are no longer confined to translator-centred decisions. They also extend to the characteristics and functioning of artificial intelligence systems, the reliability of their outputs, the conditions under which they are used, and the distribution of responsibility between human translators and technological systems.

Large language models are considered systems that, through the pre-training process carried out on large text corpora, demonstrate adaptability across various language tasks and offer a task-agnostic framework of use (Brown et al., 2020). This broad structure enables large language models to be used in the field of translation, as in many other tasks. In the modern period, machine translation systems have developed rapidly and have begun to be used in many different fields. Translation has also been directly affected by this development. Unlike earlier approaches, neural machine translation systems process translation through a single neural model. During this process, the model attends to the relevant parts of the source sentence while predicting the target word. This strengthens the connection between the source sentence and the translation output (Bahdanau et al., 2015).

As automatic translation outputs can now be produced in a very short time, evaluating their quality has become equally important. These outputs need to be assessed in terms of fluency, accuracy, coherence, and overall quality. For this reason, several automatic metrics have been developed to evaluate translation quality more quickly and systematically. Automatic evaluation metrics can broadly be divided into traditional metrics such as BLEU, ROUGE, and METEOR, which primarily rely on surface-level lexical overlap, and modern metrics such as BERTScore, LaBSE, and COMET, which focus on contextual and semantic similarity. This distinction is particularly relevant to Arabic–Turkish translation, as surface-level correspondence may remain limited in representing semantic and contextual equivalence between these morphologically complex languages (Yazal, 2026). Among the modern metrics, COMET is designed to correlate more closely with human judgments (Rei et al., 2020), while BERTScore, which is employed in the present study, measures contextual and semantic similarity (Zhang et al., 2020). Although these metrics do not directly detect hallucination in all its forms, they provide useful complementary evidence regarding the semantic proximity between translation outputs and reference translations. These developments have improved the fluency of translation outputs, but they have also brought some limitations in terms of accuracy and fidelity to the source text. Therefore, it would not be accurate to assume that these

systems always produce translations that are both correct and faithful to the source. In some cases, they may generate outputs that appear coherent and fluent, but do not sufficiently correspond to the source text or contain factually inaccurate information.

Machine translation ethics can be evaluated within the framework of reliability, non-misleading communication, clear presentation of information, and accuracy in relation to the outputs produced by automatic and artificial intelligence-based translation systems. In this regard, translation outputs should be examined in terms of accuracy and ethical responsibility. Hallucinated outputs in machine translation can be regarded as an important issue in terms of machine translation ethics, as they not only constitute a violation of equivalence in translation but also carry the risk of misleading users and distorting accurate information.

1.3. Hallucination: Definition and Classification

In the context of machine translation, hallucination is defined as the generation of content that is inconsistent with the source text or not present in the source text during the transfer of source-text information into the target language. This problem was identified with the deployment of the first neural machine translation models. Neural machine translation models are reported to tend to sacrifice adequacy for the sake of fluency, especially when evaluated on out-of-domain test sets. In this respect, the concept of hallucination is closely related to the terms faithfulness and factuality in the literature. Faithfulness is defined as the opposite of hallucination and refers to remaining consistent and truthful to the provided source. Factuality, on the other hand, refers to the quality of being actual or based on fact; however, unlike faithfulness, it may vary depending on what is accepted as the source of the "fact" (Ji et al., 2023). In this context, the definition and evaluation of hallucination may differ depending on the task under consideration and the type of information taken as reference.

In the study by Lee et al. (2018), hallucination is informally defined as a case in which the translation of a perturbed input, created by adding a word to the input sentence, has almost no words in common with the translation of the unperturbed input. Studies focusing on summarization systems show that although these systems can produce good outputs in terms of linguistic fluency, they may generate expressions that do not fully correspond to the source text or are not present in the source-text content (Maynez et al., 2020). As the concept of hallucination began to be addressed systematically, approaches to its classification were developed. In this context, drawing on Maynez et al. (2020) approach to content that is unsupported by the source input, Zhou et al. (2021, p. 1394) explain two categories of hallucination, namely intrinsic and extrinsic hallucination:

- *Intrinsic hallucination* refers to the production of information found in the source text in an incorrect, distorted, or source-inconsistent manner in the translation output.
- *Extrinsic hallucination*, on the other hand, refers to the addition of information that is not present in the source text to the translation output.
- Raunak et al. (2021) introduce several definitions to systematize the study of hallucinations in neural machine translation and classify NMT hallucinations into two primary categories:
- *Hallucinations under perturbations:* For a given source input, if the translations generated from the perturbed and unperturbed input sequences differ drastically, the model is considered to generate a hallucination under perturbation.
- *Natural hallucinations:* Natural hallucination refers to cases in which, without any perturbation applied to the source input, the generated translation is severely inadequate.
- Natural hallucinations are further divided into two subcategories:
- *Detached hallucinations:* These refer to fluent but completely inadequate translations.
- *Oscillatory hallucinations:* These refer to inadequate translations that contain repeating n-grams.

Following this classification, Guerreiro et al. (2023) take detachment from the source sequence as the main criterion for distinguishing hallucinations from other translation errors and state that largely fluent hallucinations may appear either as fully detached outputs or as strongly detached outputs that partially support the source content. The examination of hallucinations through different classifications shows that this phenomenon is not a one-dimensional type of error, thereby making the automatic analysis of such categories an important area of research. These classifications indicate that hallucinations should be addressed from a multidimensional perspective in terms of both content and model behavior. To make the hallucination classifications discussed in this section clearer and more understandable, examples are provided for each category:

Hallucination Classifications

Hallucination Type	Explanation	Source Text	Target Text	Analysis
Intrinsic Hallucination (Zhou et al., 2021)	The information contained in the source text is transferred into the target text incorrectly or in a distorted manner.	The committee concluded that the available evidence was insufficient to establish a causal relationship between the treatment and the reported adverse effects.	The committee concluded that the available evidence was sufficient to establish a causal relationship between the treatment and the reported adverse effects.	The source text states that the available evidence was insufficient to establish a causal relationship, whereas the target text states that it was sufficient. This reversal contradicts information explicitly contained in the source text. Therefore, it constitutes an example of intrinsic hallucination under the classification adopted in this study.
Extrinsic Hallucination (Zhou et al., 2021)	The inclusion of additional information or content in the translation output that has no explicit basis in the source text.	The archival documents indicate that the agreement was signed by the two countries in 1978.	The archival documents indicate that the agreement was signed by the two countries in 1978 following five years of secret negotiations.	The target text preserves the source information concerning the parties and the date of the agreement. However, it introduces the claim that the agreement followed five years of secret negotiations, although this information has no basis in the source text. This unsupported addition therefore constitutes an example of extrinsic hallucination.
Detached Hallucination (Raunak et al., 2021)	A case in which the translation is fluent but completely inadequate.	The prolonged drought caused grain production in the region to fall to its lowest level in a decade.	The new agricultural support program significantly improved local farmers' access to international markets.	Although the target text is fluent and remains within the general domain of agriculture, it does not convey the source information concerning the prolonged drought or the decline in grain production. Instead, it generates unrelated content about an agricultural support program and access to international markets. The fluent but severely inadequate

				output is therefore substantially detached from the source content and constitutes an example of detached hallucination.
Oscillatory Hallucination (Raunak et al., 2021)	A case in which an inadequate translation containing repeated n-gram structures occurs in the translation output.	The delegation withdrew the proposal after the negotiations had failed.	The delegation withdrew the proposal after the negotiations had failed, after the negotiations had failed, after the negotiations had failed.	The target text conveys the main source information but repeatedly generates the n-gram sequence “after the negotiations had failed” without any corresponding repetition in the source text. This repetitive generation pattern reduces the adequacy and coherence of the output and constitutes an example of oscillatory hallucination.
Fully Detached Hallucination (Guerreiro et al., 2023)	A case in which the target output has no semantic relation to the source text.	The court ruled that evidence obtained unlawfully could not be admitted in the proceedings.	Recent observations indicate that glaciers in the Southern Hemisphere are melting faster than previously estimated.	Although the target text is fluent and grammatically well-formed, it has no semantic relation to the source text. It replaces the legal information concerning the admissibility of evidence with an unrelated statement about glacier melting. This complete disconnection from the source content therefore constitutes an example of fully detached hallucination.
Strongly Detached Hallucination (Guerreiro et al., 2023)	A case in which the translation output contains a limited element related to the source text but largely moves away from the source content.	Several pottery fragments dating from the Roman period were discovered during the archaeological excavations.	Pottery fragments dating from the Roman period were discovered during the archaeological excavations, conclusively proving that the settlement was the regional capital of the empire and accommodated approximately fifty thousand soldiers.	The target text retains the source information concerning the discovery of Roman-period pottery fragments. However, it moves substantially beyond the source by claiming that the settlement was the regional capital of the empire and accommodated approximately fifty thousand soldiers. Since only a limited part of the output is supported by the source, while the remaining content consists of extensive unsupported claims, the output constitutes an example of strongly detached hallucination.

Source. Prepared by the author.

These classifications may seem close to one another at some points, but they do not focus on the same aspect of hallucination. Zhou et al. (2021) make a distinction based on the source of unsupported content: it may result either from the distortion of information in the source text or from the addition of information that is not included in the source. Raunak et al. (2021) discuss hallucinations in neural machine translation through perturbation-based and naturally occurring cases. Guerreiro et al. (2023), however, approach the issue by looking at how far the output moves away from the source sequence. For this reason, these classifications should not be treated as entirely separate or opposing categories. They are better understood as complementary perspectives that explain different sides of hallucination.

From the perspective of translation studies, hallucination is more than a technical error. It also disrupts the equivalence relationship between the source and target texts. When this relationship is weakened, problems may appear in meaning, content, and function. This makes hallucination an important issue in the translation process. Intrinsic hallucinations may harm accuracy and reliability because they involve an incorrect transfer of source meaning or content. Extrinsic hallucinations may weaken fidelity and ethical reliability because they add information that does not exist in the source text. Detached and oscillatory hallucinations may also reduce the acceptability of translation by damaging the integrity and communicative function of the source text. In this sense, hallucination can be seen as a multidimensional translation problem related to reliability, equivalence, accuracy, and acceptability.

The effects of hallucination may also change according to text type. In medical texts, hallucinated translations may create risks that directly affect human health. In commercial or diplomatic texts, they may cause misunderstandings and communication problems. This shows that hallucination has a particularly important place in translation, especially in texts where accuracy and reliability are necessary. In practice, automatic translation outputs cannot always be checked continuously and in detail. For this reason, users should verify these outputs, especially when the content is medical, legal, or official. Translators also play an important role in reducing hallucination-related risks through proofreading and revision. In addition, users should be careful when sharing personal or sensitive information with data-driven systems, since these systems may produce inaccurate or hallucinated outputs.

Methods

This study follows a qualitative research design to analyze hallucination types in Turkish–Arabic machine translation outputs. The analysis is carried out on three Turkish source texts from different text types: literary, legal, and medical/health texts. These texts were chosen because they include linguistic, terminological, and field-related features that can make hallucinated deviations more visible in translation. For each Turkish source text, an Arabic human translation was prepared and used as the reference translation. The same source texts were then translated into Arabic by ChatGPT 5.5, Gemini 3.5 Flash, and Grok 4.3. When obtaining translation outputs from the large language models, the same prompt was given to all systems, and they were asked only to translate the source text into Arabic. Accordingly, the prompt was standardized as "Translate only this text into Arabic". The outputs produced by these systems were compared with the reference translations in order to identify hallucinated deviations. Therefore, the study is not based on researcher-created artificial examples, but on actual translation outputs generated by large language models.

The analysis focused on fidelity to the source text, semantic integrity, terminological accuracy, source-external additions, task-irrelevant explanations, semantic narrowing, and repetition. The literary text was examined in terms of interpretive expansion and task-irrelevant explanation; the legal text in terms of conceptual accuracy, terminological consistency, semantic narrowing, and

source-external addition; and the medical/health text in terms of medical terminology and domain-specific meaning transfer. In addition to qualitative analysis, BERTScore was used as a complementary indicator to measure semantic similarity between the system outputs and the reference translations. The BERTScore F1 value ranges between 0 and 1, and a value closer to 1 indicates a higher level of semantic similarity between the candidate translation and the reference translation. The BERTScore calculations were carried out in a Python-based Google Colaboratory environment.

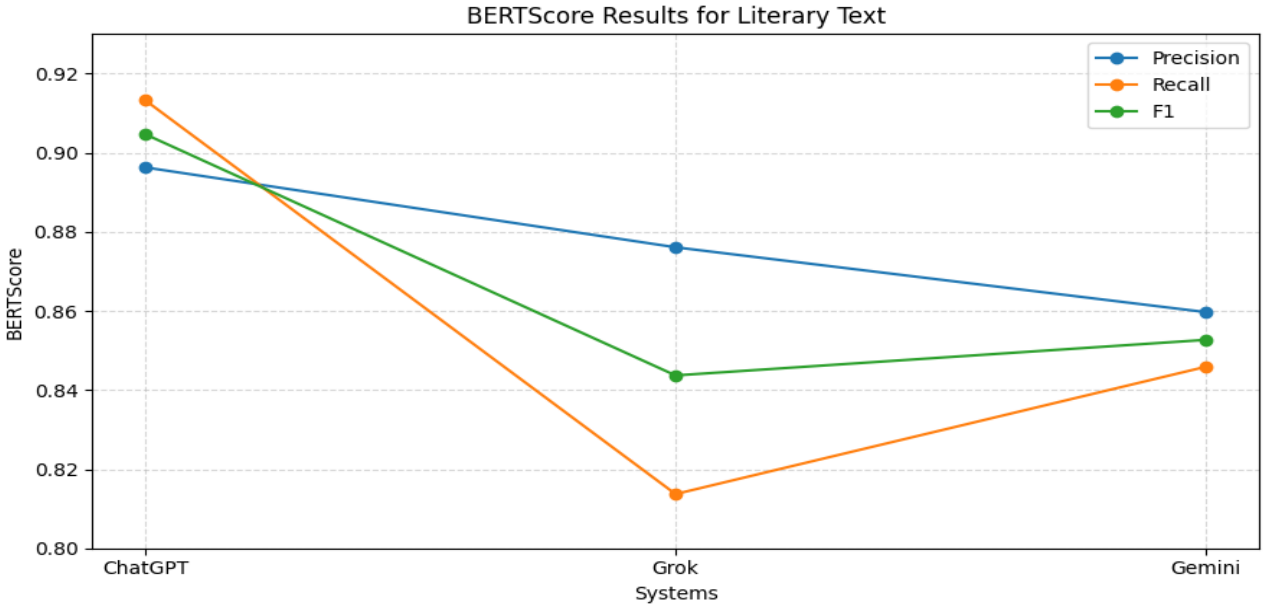
2. Findings

1.4. Hallucination Example in Literary Translation
Source Text:
İkimiz de herkes gibiyiz.
Neden kendi bakışlarını bırakıp yapmacık gözlerle bakar?
Neden çerçevesini bozarsın dudaklarının?
Allı pullu tavırlara kim kanar? (Asaf, 2025, p. 26).
English Translation: We are both like everyone else.
Why do you abandon your own gaze and look through affected eyes?
Why do you distort the outline of your lips?
Who is taken in by such ostentatious airs?
Human Translation:

نَحْنُ الاثْنَانِ كَسَائِرِ النَّاسِ.
لِمَ تَتْرُكُ نَظَرَاتِكَ وَتَنْظُرُ بِعُيُونٍ مُتَكَلَّفَةٍ؟
لِمَ تُفْسِدُ إِطَارَ شَفَتَيْكَ؟
مَنْ يَنْخَدِعُ بِتَصَرُّفَاتٍ مُزَخْرَفَةٍ؟

System	Translation Output	BERTScore F1	Hallucination / Deviation Type
ChatGPT	نَحْنُ، أَنَا وَأَنْتَ، لا نَخْتَلِفُ عَنْ سَائِرِ البَشَرِ، فَلِمَ تَهْجُرُ نَظَرَاتِكَ الحَقِيقِيَّةَ، وَتَسْتَعِيرُ عَيْنَيْنِ مِنْ زِيفٍ؟ لِمَ تُبَعْثِرُ رَسْمَ شَفَتَيْكَ؟ وَمَنْ ذَا الَّذِي يَغْتَرُّ بِحَرَكَاتٍ مُزَيَّنَةٍ، لَامِعَةٍ كَذِبَةٍ؟ *English translation: We–you and I–are no different from the rest of humankind. Why, then, do you abandon your genuine gaze and borrow two eyes of falseness? Why do you scatter the shape of your lips? And who could be deceived by adorned gestures–glittering and false?*	0.9047	Interpretive expansion / source-external addition
Grok	نحن الاثنان كالناس جميعاً. فلمَ تترك نظرتكَ الحقيقية، وتنظر بعيونٍ مصطنعةٍ مزيّفة؟ ولمَ تُفسد إطارَ شفتيكِ؟ من ينخدعُ بالألوان الزاهية والحركاتِ المُزخرفة؟ كلامٌ بيننا... نحن الاثنان كالناس جميعاً *English translation: We are both like everyone else. Why do you abandon your genuine gaze and look through artificial, false eyes? Why do you distort the outline of your lips? Who is deceived by the bright colours and ornate gestures? Words between us ... We are both like everyone else.*	0.8437	Extrinsic hallucination / repetition / partial semantic expansion
Gemini	إليك ترجمة هذه الكلمات بنبرة شعرية تحاكي روح النص الأصلي: كلانا كبقية الناس، عادِيَّان تماماً... فلِمَ يركُدُ المرءُ عن نظراتهِ	0.8527	Task-irrelevant explanation / interpretive

System	Translation Output	BERTScore F1	Hallucination / Deviation Type
	الخاصّة، لِيَنظُرَ بأعيُنٍ مُستعارةٍ وزائفة؟ ولِمَ تُشوّهُ الإطارَ الطبيعيَّ لِشفتيك؟ ومَن ذا الذي تُغويهِ تلك الأطوارُ المُزركشة؟ *English translation: Here is a translation of these words in a poetic tone that emulates the spirit of the original text: We are both like everyone else, entirely ordinary ... Why does one hold back from one's own gaze to look through borrowed and false eyes? Why do you distort the natural outline of your lips? And who could be enticed by such ostentatious mannerisms?*		addition / partial semantic shift



In this example, when the BERTScore F1 values, that is, the harmonic mean of precision and recall, are examined, it is observed that ChatGPT 5.5 produced the output that is semantically closest to the human translation, with a score of 0.9047. However, the qualitative analysis shows that the ChatGPT output contains interpretive expansions and source-external aesthetic additions that are not present in the source text. Expressions such as "تَسْتَعِيرُ عَيْنَيْنِ مِنْ زِيفٍ" and "لَامِعَةٍ كَذِبَةٍ"are not directly found in the source text; rather, they appear in the target text as interpretive elements that expand the poetic meaning. Therefore, although the BERTScore value is high, the presence of source-external additions shows that BERTScore may remain limited when used alone to detect hallucinated elements.

Grok 4.3, on the other hand, received the lowest score in this example, with a BERTScore F1 value of 0.8437. The expression "كلامٌ بيننا..." in the output constitutes an addition that is not present in the source text. In addition, the repetition of the first sentence at the end of the output changes the structural organization of the source text. These findings show that the Grok output contains elements of extrinsic hallucination, repetition, and partial semantic expansion. Nevertheless, the fact that the score is still relatively close to 1 indicates that BERTScore can show contextual/semantic proximity, but cannot fully capture hallucinated deviations such as repetition and source-external addition.

Gemini 3.5 Flash produced a result close to Grok, with a BERTScore F1 value of 0.8527. Unlike the other systems, the Gemini output begins with the explanatory sentence " إليك ترجمة هذه الكلمات بنبرة شعرية تحاكي روح النص الأصلي", which is presented as if it were part of the translation output. Since this

sentence does not correspond to any element in the source text and does not belong to the target translation itself, it is evaluated as a task-irrelevant explanation. This indicates that the model moves beyond the expected translation task by inserting a meta-textual comment into the output. In addition, expressions such as "الإطار الطبيعي" and "أعين مستعارة" stand out as interpretive additions that are not directly present in the source text.

In this context, the findings show that BERTScore is useful for indicating semantic proximity between automatic system outputs and human translation. However, it may remain limited on its own in detecting hallucinated deviations in literary texts, such as source-external addition, interpretive expansion, repetition, and task-irrelevant explanation. Therefore, in this study, BERTScore was not used as a tool that directly measures hallucination, but as a complementary indicator supporting qualitative translation analysis.

1.5. Hallucination Example in Legal Translation

Source Text:

Tacirler arasında, diğer tarafı temerrüde düşürmeye, sözleşmeyi feshe, sözleşmeden dönmeye ilişkin ihbarlar veya ihtarlar noter aracılığıyla, taahhütlü mektupla, telgrafla veya güvenli elektronik imza kullanılarak kayıtlı elektronik posta sistemi ile yapılır (Türk Ticaret Kanunu, 2011).

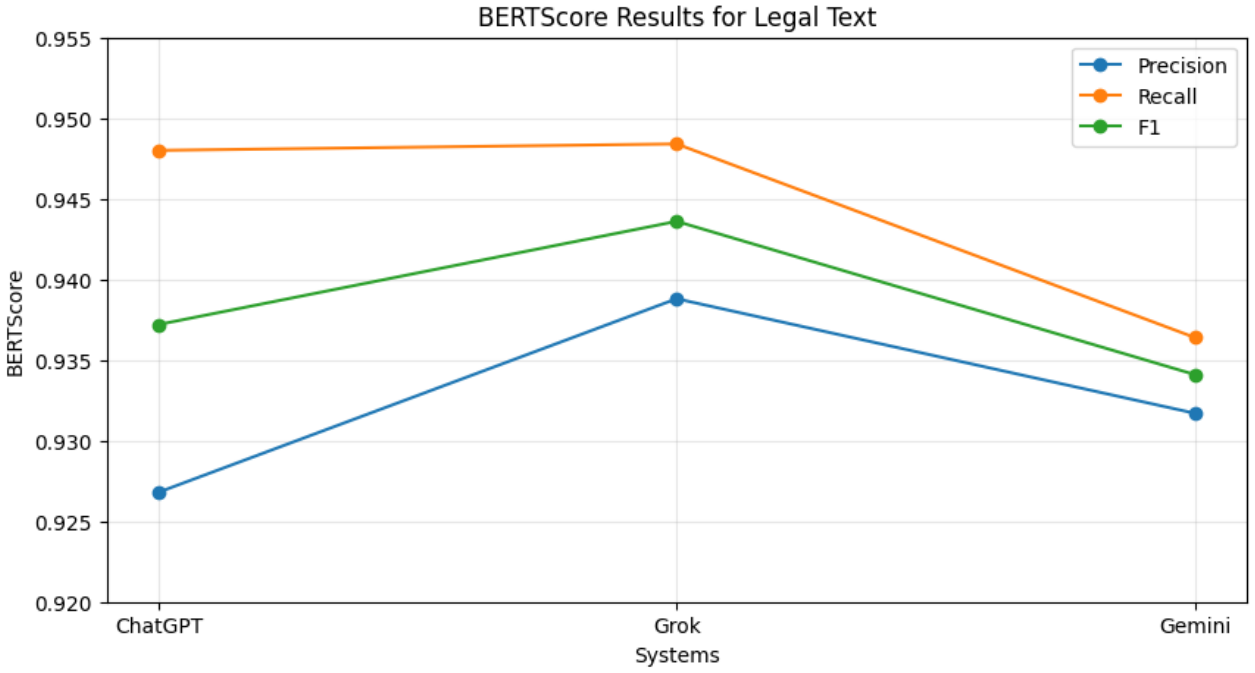
English Translation: Between merchants, notices or warnings relating to placing the other party in default, terminating the contract, or rescinding it shall be served through a notary public, by registered letter, by telegram, or through the registered electronic mail system using a secure electronic signature.

Human Translation:

تُجرى الإخطارات أو الإنذارات بين التجار، المتعلقة بوضع الطرف الآخر في حالة تخلّف عن الوفاء أو بفسخ العقد أو بالرجوع عنه، بواسطة كاتب العدل أو برسالة مسجلة أو بالبرقية أو عبر نظام البريد الإلكتروني المسجل باستخدام التوقيع الإلكتروني الآمن.

System	Translation Output	BERTScore F1	Hallucination / Deviation Type
ChatGPT	تُوجَّه الإخطارات أو الإنذارات المتبادلة بين التجّار، والمتعلقة بإعذار الطرف الآخر وجعله في حالة تخلّف عن الوفاء، أو بفسخ العقد، أو بالرجوع عنه، بواسطة الكاتب بالعدل/الموثّق، أو بموجب كتابٍ مسجّل، أو برقية، أو عن طريق نظام البريد الإلكتروني المسجّل باستخدام التوقيع الإلكتروني الآمن. *English translation: Notices or warnings exchanged between merchants concerning formally notifying the other party and placing them in a state of failure to perform, terminating the contract, or withdrawing from it shall be delivered through a notary public/notary, by registered letter, by telegram, or through the registered electronic mail system using a secure electronic signature*.	0.9372	No clear hallucination / limited explanatory expansion
Grok	بين التجار، تتم الإشعارات أو الإنذارات المتعلقة بإيقاع الطرف الآخر في التخلف عن الوفاء (التمرّد)، أو فسخ العقد، أو الرجوع عن العقد، عن طريق الكاتب العدل (الموثق)، أو بالخطاب المسجل، أو بالبرقية، أو عن طريق نظام البريد الإلكتروني المسجل باستخدام التوقيع الإلكتروني الآمن. *English translation: Between merchants, notices or warnings concerning placing the other party in a state of failure to perform (rebellion), terminating*	0.9436	Intrinsic hallucination / terminological deviation / source-external addition

System	Translation Output	BERTScore F1	Hallucination / Deviation Type
Gemini	*the contract, or withdrawing from the contract shall be delivered through a notary public (notary), by registered letter, by telegram, or through the registered electronic mail system using a secure electronic signature.* تتم الإخطارات أو الإنذارات المتبادلة بين التجار، والمتعلقة بإعذار الطرف الآخر (وضعه في حالة التخلف عن السداد)، أو فسخ العقد، أو العدول عنه، عن طريق الكاتب بالعدل، أو الخطاب المسجل بعلم الوصول، أو البرق، أو عبر نظام البريد الإلكتروني المسجل باستخدام التوقيع الإلكتروني الآمن. *English translation: Notices or warnings exchanged between merchants concerning formally notifying the other party (placing them in a state of failure to make payment), terminating the contract, or withdrawing from it shall be delivered through a notary public, by registered letter with acknowledgment of receipt, by telegraph, or through the registered electronic mail system using a secure electronic signature.*	0.9341	Extrinsic hallucination / intrinsic semantic narrowing



In this example, when the BERTScore F1 values are examined, it is observed that Grok 4.3 produced the output that is semantically closest to the human translation, with a score of 0.9436. ChatGPT 5.5 and Gemini 3.5 Flash produced results that are close to each other and to Grok 4.3. However, the qualitative analysis shows that these scores do not, on their own, indicate whether legal terms are transferred in a reliable, accurate, and hallucination-free manner.

When the ChatGPT 5.5 output is examined, it is observed that the general meaning of the source text is largely preserved. The expression "إعذار الطرف الآخر" constitutes an appropriate legal equivalent for placing the other party in default. However, the additional phrase "وجعله في حالة تخلّف عن الوفاء" restates the concept of default in an explanatory manner. Since this phrase does not introduce information

that contradicts or departs from the source meaning, it should not be regarded as a source-external hallucination. Therefore, although the ChatGPT 5.5 output does not constitute a clear case of hallucination, it can be evaluated as a limited explanatory expansion. The Grok 4.3 output, despite having the highest BERTScore F1 value, constitutes a hallucinated example in terms of qualitative analysis. Although the expression "التخلف عن الوفاء" broadly conveys the legal concept of default, the parenthetical addition "التمرّد" introduces an unsupported and legally inaccurate gloss. Since this addition shifts the source concept into a different semantic field and evokes meanings such as rebellion, revolt, or resistance, it constitutes an intrinsic hallucination, a terminological deviation, and a source-external addition.

In the Gemini 3.5 Flash output, the expression "التخلف عن السداد" limits the concept of default only to the non-payment of a debt. However, the concept of "temerrüt" in the source text refers to a broader legal context. Therefore, this expression leads to intrinsic semantic narrowing. In addition, the expression "بعلم الوصول" was added to the target text as an element not present in the source text and thus constitutes a source-external addition.

1.6. Hallucination Example in Medical Translation

Source Text:

A grubu beta-hemolitik streptokoklara bağlı tonsillofarenjite ikincil olarak gelişen yaygın enflamatuvar bir bağ dokusu hastalığıdır. Streptokok antijenlerine karşı gelişen antikorların eklem, kalp, bazal gangliyon gibi dokularda oluşturduğu hasar sonucu gelişir (TC Sağlık Bakanlığı, 2003, p. 23).

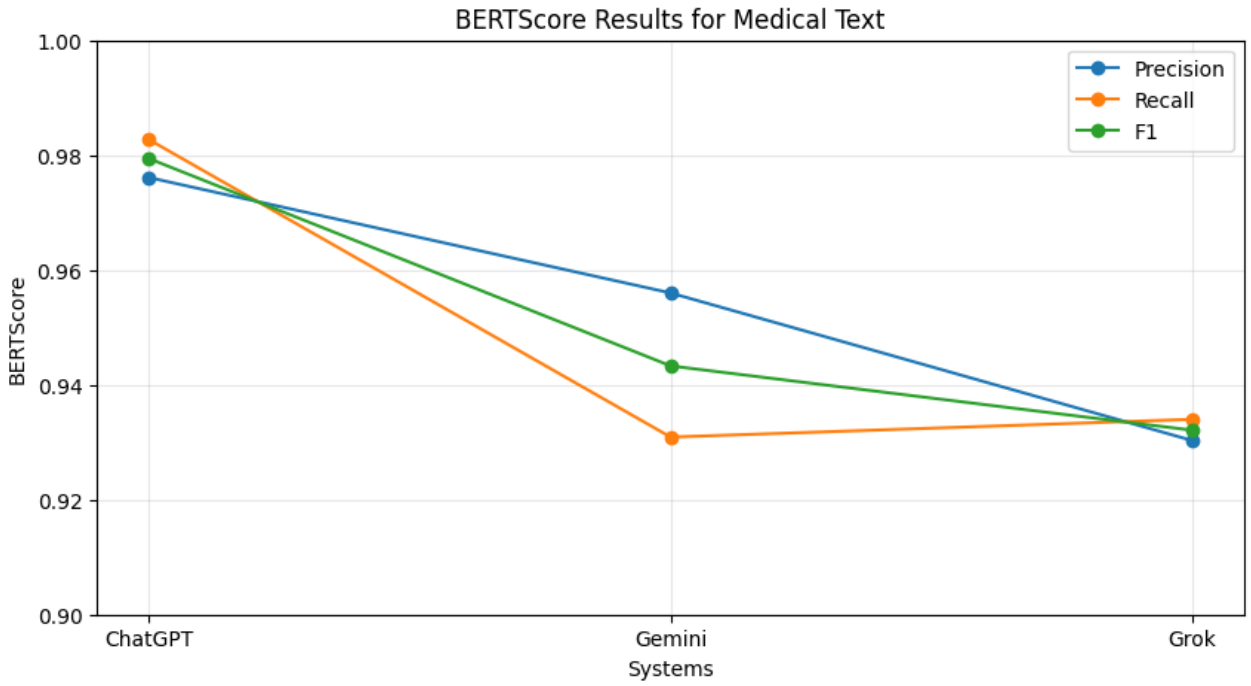
English Translation: It is a widespread inflammatory disease of the connective tissue that develops secondary to tonsillopharyngitis caused by group A beta-hemolytic streptococci. It results from damage caused by antibodies formed against streptococcal antigens in tissues such as the joints, heart, and basal ganglia.

Human Translation:

هو مرضٌ التهابيٌّ منتشر يصيب النسيج الضام، ويتطوّر ثانويًّا لالتهاب اللوزتين والبلعوم الناجم عن العقديات بيتا الحالّة للدم من المجموعة .A وينشأ نتيجة الضرر الذي تُحدثه الأجسام المضادة المتكوّنة ضد مستضدات العقديات في أنسجة مثل المفاصل والقلب والعقد القاعدية.

System	Translation Output	BERTScore F1	Hallucination / Deviation Type
ChatGPT	هو مرضٌ التهابيٌّ واسع الانتشار يصيب النسيج الضام، ويتطوّر ثانويًّا لالتهاب اللوزتين والبلعوم الناجم عن المكورات العقدية بيتا الحالة للدم من المجموعة .A وينشأ نتيجة الضرر الذي تُحدثه الأجسام المضادة المتكوّنة ضد مستضدات المكورات العقدية في أنسجة مثل المفاصل والقلب والعقد القاعدية. *English translation: It is a widespread inflammatory disease affecting the connective tissue and develops secondary to tonsillopharyngitis caused by group A beta-hemolytic streptococci. It arises as a result of the damage caused by antibodies formed against streptococcal antigens in tissues such as the joints, heart, and basal ganglia.*	0.9794	No clear hallucination
Gemini	هو مرض التهابي واسع النطاق يصيب النسيج الضام، ويتطور كعامل ثانوي لالتهاب اللوزتين والبلعوم الناجم عن المكورات العقدية الحالة للدم من المجموعة (أ). ويحدث نتيجة للأضرار التي تسببها	0.9433	Limited intrinsic semantic shift

System	Translation Output	BERTScore F1	Hallucination / Deviation Type
Grok	الأجسام المضادة المتكونة ضد مستضدات المكورات العقدية في أنسجة مثل المفاصل والقلب والعقد القاعدية. *English translation: It is a wide-ranging inflammatory disease affecting the connective tissue and develops as a secondary factor to tonsillopharyngitis caused by group A hemolytic streptococci. It occurs as a result of the damage caused by antibodies formed against streptococcal antigens in tissues such as the joints, heart, and basal ganglia.* الترجمة إلى العربية: هو مرض منتشر في النسيج الضام الالتهاب يتطور ثانويًا لالتهاب اللوزتين والبلعوم الناتج عن المكورات العقدية بيتا-هيموليتيكية من المجموعة A. ينشأ نتيجة الضرر الذي تسببه الأجسام المضادة المتكونة ضد مستضدات العقديات في أنسجة مثل المفاصل والقلب والبنوى القاعدية (البازال غانغليون). *English translation: Translation into Arabic: It is a widespread disease in the inflammatory connective tissue that develops secondary to tonsillopharyngitis caused by group A beta-hemolytic streptococci. It arises as a result of the damage caused by antibodies formed against streptococcal antigens in tissues such as the joints, heart, and the basal nuclei (the basal ganglion).*	0.9321	Task-irrelevant explanation / terminological deviation / limited intrinsic semantic shift

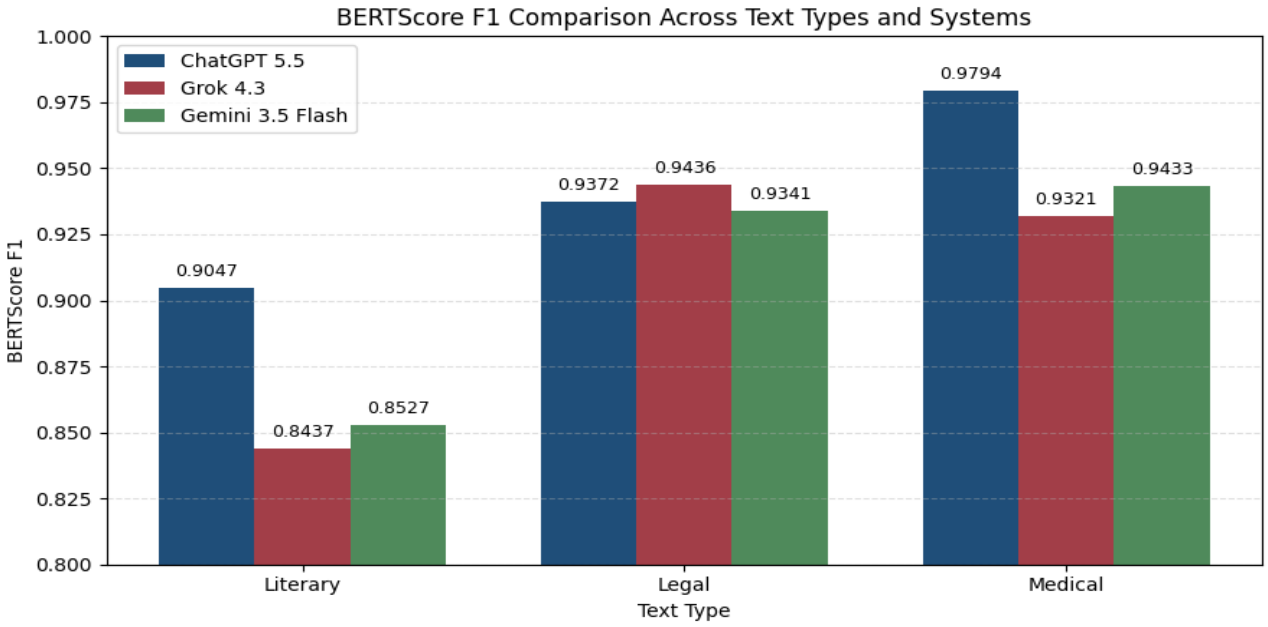


When the BERTScore F1 scores are evaluated, it is observed that ChatGPT 5.5 produced the output that is semantically closest to the human translation, receiving the highest score of 0.9794. Gemini 3.5 Flash, with a score of 0.9433, and Grok 4.3, with a score of 0.9321, received lower scores than ChatGPT 5.5. In this example, the quantitative findings and qualitative analysis results largely

overlap. Nevertheless, qualitative analysis remains necessary for identifying deviations involving medical terminology and domain-specific meaning. When the ChatGPT 5.5 output is examined, it is observed that the source meaning and the medical relationship between tonsillopharyngitis, group A beta-hemolytic streptococci, antibodies, and tissue damage are accurately preserved. The medical terms are also rendered appropriately, and no source-external information or task-irrelevant explanation is introduced. Therefore, no clear hallucination or deviation type was identified in the ChatGPT 5.5 output. In the Gemini 3.5 Flash output, the expression "يتطور كعامل ثانوي" presents the disease as developing "as a secondary factor", whereas the source text states that it develops secondary to tonsillopharyngitis. This formulation partially alters the causal relationship expressed in the source text. Since the general medical meaning is preserved but this relationship is rendered less accurately, the output can be evaluated as containing a limited intrinsic semantic shift.

In the Grok 4.3 output, the phrase "الترجمة إلى العربية" :constitutes a task-irrelevant explanation because it is not part of the source text or the requested translation. The formulation " مرض منتشر في النسيج الضام الالتهابي" also partially shifts the source meaning by presenting the condition as a widespread disease “in the inflammatory connective tissue”, rather than as a widespread inflammatory disease of the connective tissue. In addition, the rendering "البنوى القاعدية (البازال غانغليون)" constitutes a problematic and non-standard formulation of the medical concept in this context. Therefore, the Grok output contains a task-irrelevant explanation, a terminological deviation, and a limited intrinsic semantic shift.

1.7. General Evaluation



When the findings obtained from the literary, legal, and medical text examples are evaluated together, it is seen that hallucinated deviations in the translation outputs of large language models vary according to text type. BERTScore F1 results provide a useful quantitative indicator for showing the general semantic proximity of system outputs to human translations. However, comparison of the BERTScore results with the qualitative findings reveals that high similarity values do not always indicate fidelity to the source text, terminological accuracy, or translation reliability. While deviations in literary texts appear mainly in the form of interpretive expansion, source-external aesthetic addition, repetition, and task-irrelevant explanation, legal texts foreground terminological deviation, conceptual narrowing, and source-external addition. In the medical text example, BERTScore results and qualitative analysis findings overlap to a greater extent; nevertheless, it is

observed that terminologically sensitive uses still need to be evaluated through human expertise. Accordingly, the findings show that automatic metrics are not sufficient on their own in the evaluation of translation hallucinations and should be used as complementary elements alongside qualitative analysis from a translation studies perspective.

Conclusion

This article analyzed hallucination in Turkish–Arabic machine translation outputs from a translation studies perspective and focused on how this issue appears in literary, legal, and medical texts. The outputs produced by ChatGPT 5.5, Grok 4.3, and Gemini 3.5 Flash were compared with Arabic human translations. They were evaluated according to fidelity to the source text, semantic equivalence, terminological accuracy, source-external addition or omission, task-irrelevant explanation, and repetition. The findings show that hallucination does not take the same form in every text type. Instead, it changes according to the nature and function of the text. In literary texts, deviations were mostly seen as interpretive expansion, source-external aesthetic addition, task-irrelevant explanation, and repetition. In legal texts, terminological/conceptual deviation, semantic narrowing, and source-external addition became more visible. In medical texts, the qualitative and quantitative findings were more closely aligned than in the other two text types. However, texts that require terminological precision still cannot be evaluated only through automatic metrics. These findings show that hallucination is not only a technical or linguistic error, but also a multidimensional transfer problem related to source-text fidelity, semantic equivalence, terminological accuracy, reliability, and ethical responsibility.

The study also shows that high BERTScore values do not always indicate full fidelity to the source text, terminological accuracy, or the absence of hallucination. Although BERTScore is a useful supporting tool for indicating the general semantic proximity between the reference translation and the system output, it remains limited in detecting hallucination-related problems such as the addition of information not present in the source text, terminological deviation, task-irrelevant explanation, repetition, and semantic narrowing on its own. In this context, hallucination analysis in machine translation should combine automatic evaluation metrics with qualitative source-based analysis from a translation studies perspective. This study was conducted on three examples in order to demonstrate different manifestations of hallucination in Turkish–Arabic machine translation. Future studies may conduct multidimensional analyses supported by larger datasets, different text types, different machine translation systems, and human evaluators; different automatic evaluation metrics may also be used comparatively. In this respect, the study contributes to the discussion of hallucination in Turkish–Arabic machine translation from a translation studies perspective.

References

Asaf, Ö. (2025). *Lavinia: Aşk şiirleri*. Yapı Kredi Yayınları.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.1409.0473

Baker, M. (2018). *In other words: A coursebook on translation* (3rd ed.). Routledge. https://doi.org/10.4324/9781315619187

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Guerreiro, N. M., Voita, E., & Martins, A. (2023). Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 1059–1075). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.eacl-main.75

Gürçağlar, Ş. (2016). *Çevirinin ABC'si*. Say Yayınları.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), Article 42, 1–55. https://doi.org/10.1145/3703155

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), Article 248, 1–38. https://doi.org/10.1145/3571730

Kimera, R., Kim, Y.-S., & Choi, H. (2024). *Advancing AI with integrity: Ethical challenges and solutions in neural machine translation*. arXiv. https://doi.org/10.48550/arXiv.2404.01070

Lee, K., Firat, O., Agarwal, A., Fannjiang, C., & Sussillo, D. (2018). *Hallucinations in neural machine translation*. OpenReview.

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173

Nida, E. A. (2000). Principles of correspondence. In L. Venuti (Ed.), *The translation studies reader* (126–140). Routledge.

Raunak, V., Menezes, A., & Junczys-Dowmunt, M. (2021). The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1172–1183). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.92

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2685– 2702. https://doi.org/10.18653/v1/2020.emnlp-main.213

TC Sağlık Bakanlığı. (2003). *Birinci basamağa yönelik tanı ve tedavi rehberleri* (2. bs.).

Toury, G. (1980). *In search of a theory of translation*. The Porter Institute for Poetics and Semiotics.

Türk Ticaret Kanunu. (2011, 14 şubat). *T.C. Resmî Gazete* (27846), Kanun No. 6102.

Yazal, Y. (2026). BLEU-dan COMET-ə: Ərəb–türk audiovizual tərcüməsində ənənəvi və müasir qiymətləndirmə metriklərinin müqayisəli təhlili. *Qədim Diyar beynəlxalq elmi jurnal, 8*(6), 103–120. https://doi.org/10.36719/2706-6185/61/103-120

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*. https://openreview.net/forum?id=SkeHuCVFDr

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2025). Siren's song in the AI ocean: A survey on hallucination in large language models. *Computational Linguistics*, *51*(4), 1373–1418. https://doi.org/10.1162/coli.a.16

Zhou, C., Neubig, G., Gu, J., Diab, M., Guzmán, F., Zettlemoyer, L., & Ghazvininejad, M. (2021). Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (1393–1404). https://doi.org/10.18653/v1/2021.findings-acl.120