Assessing the Accuracy of AI-Generated Idiom Translations

Marijana Gašparović
Faculty of Humanities and Social Sciences,
University of Rijeka
Croatia
mgasparovic99@gmail.com

Marija Brala Vukanović
Faculty of Humanities and Social Sciences,
University of Rijeka
Croatia
mbrala@ffri.hr

Marija Brkić Bakarić
Faculty of Informatics and Digital Technologies,
University of Rijeka
Croatia
mbrkic@uniri.hr

Abstract

Idioms pose unique challenges for machine translation due to their metaphorical nature and cultural nuances. Consequently, they often present a translation problem even for humans. This longitudinal study evaluates the performance of ChatGPT in translating idiomatic expressions between English and Croatian, comparing results across two time points. The test set comprises 72 idioms in each translation direction, divided into three categories based on equivalence: complete, partial, and zero, with each category representing one-third of the set. The evaluation considers three layers: translation of the isolated idiom, translation of an online excerpt containing the idiom, and translation of a self-constructed example sentence. As expected, accuracy generally declined with decreasing equivalence. However, a follow-up study conducted six months later highlighted the need for continuous monitoring of machine translation tools.

1 Introduction

The impact of artificial intelligence (AI) on the language industry is evident in last year's European Language Industry Survey (ELIS), which has been conducted annually since 2013. The 2024 edition integrated and captured perceptions, expectations, and realities of AI across various segments of the industry (ELIS, 2024). Findings from the most recent survey indicate that machine translation (MT) is now used in over 50% of professional translation tasks (ELIS, 2025). Furthermore, language service providers anticipate a continued decline in both their own activity and the global language industry as a whole.

A notable trend has emerged in the MT landscape. While DeepL continues to dominate the rankings, generative AI tools such as ChatGPT are approaching—and may soon surpass—Google Translate in terms of usage. However, the growing dependence on these tools is a double-edged sword. According to ELIS (2025), AI's increased dominance has led to greater polarization within the industry. Both independent professionals and language companies attribute this to indiscriminate client use, resulting in quality degradation, diminished appreciation of linguistic expertise, and intensified price pressure.

MT still faces serious quality issues, and the accuracy of the translation heavily depends on human review. This is particularly evident in the translation of idioms and other culturally nuanced and/or contextually embedded meanings. Identifying an idiom, understanding its meaning, and finding an appropriate equivalent in the target language is a complex task that cannot be easily automated. The difficulty arises both from differences in conceptual grounding across languages and from structural divergences between them. Not all idioms exist in every language, and sometimes corresponding meanings are cross-linguistically rendered by non-corresponding linguistic forms. Since idiomatic phrases typically cannot be translated literally, achieving an adequate cross-linguistic and cross-cultural match requires deep and thorough familiarity with the idiomatic expressions of both the source and target languages, as well as their respective cultures.

Baker (1992) highlights three key challenges specific to idiom translation:

- identifying idiomatic expressions,
- interpreting their meaning, and
- accurately conveying their nuanced meanings in the target language.

She identifies five actions to avoid when translating idioms: omission, addition, word replacement, mo-

Proceedings of the Workshop on Beyond English: NLP for all Languages in an Era of LLM from Texts associated with RANLP 2025,
pages 131–140, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-105-9-015

dification of word order and changes in the grammatical structure.

Adelnia and Dastjerdi (2011) outline four strategies for translating idioms: (1) using an idiom equivalent in both meaning and form, (2) using an idiom equivalent in meaning but not form, (3) paraphrasing, and (4) omitting the idiom altogether. While full equivalence in both meaning and form is rare, paraphrasing or substituting the original idiom with a semantically equivalent expression in the target language remains the most commonly applied approach.

Previous studies involving the languages examined in this research have shown that Google Translate predominantly produced literal translations of idioms, particularly when translating from English into Croatian (Manojlović et al., 2017). Baziotis et al. (2022) noted that research on idioms in neural machine translation (NMT) remains limited, while Li et al. (2024) emphasized the particular challenges idioms pose for Transformer-based systems. While Zhu et al. (2024) demonstrated that LLMs outperformed other state-of-the-art models, Donthi et al. (2025) found that GPT-4 outperformed GPT-3.5-Turbo in translating idioms.

The aim of this paper is twofold. First, we seek to evaluate the accuracy of idiom translation. For this purpose, we adopt the classification proposed by Barchudarow (1979) as cited in Gläser (1984), which recognizes three categories of idioms based on their equivalence level in translation: complete, partial, and zero equivalence. Complete equivalence implies correspondence in both structure and meaning (e.g., to have one's head in the clouds and "biti glavom u oblacima" – which is a literal translation of the English idiom into Croatian), partial equivalence suggests that the idioms align in either structure or meaning (e.g., wear one's heart on one's sleeve and "nositi srce na dlanu" – lit. wear one's heart on the palm), but not both, and zero equivalence occurs when no similar expression exists in the target language (e.g., to hold one's horses and "stati na loptu" – lit. step on the ball).

Second, we aim to investigate whether the quality of idiom translation using a generative AI service improves over time, as might be expected.

The remainder of this paper is organized as follows. The next section outlines the research design and evaluation process. Results are presented in the third section, followed by a discussion, a summary of the main findings, and suggestions for future work. The paper is concluded with ethical considerations and limitations of the current study.

2 Methodology

The aim of this study was twofold. First, we sought to assess the accuracy of idiom translations produced by ChatGPT, currently the most widely used generative AI service in the language industry (ELIS, 2025). Second, we aimed to determine whether the quality of idiom translation using a generative AI service improves over time.

2.1 Dataset

Three lists of idioms were compiled for this study, with each idiom examined from a cross-linguistic perspective both in isolation and within context. Contextualized examples were drawn from the web. Given that the exact contents of the GPT training corpus are not publicly available, additional examples were constructed by the author (AOC) to ensure unbiased evaluation. Idioms were categorized according to their level of equivalence—complete, partial, or zero equivalence—between English and Croatian.

The dataset comprised 24 idioms for each equivalence level and translation direction, yielding a balanced distribution across categories and a total of 72 idioms per translation direction.

2.2 Method

The research was conducted using the free tier of ChatGPT. The initial assessment took place in May 2024, followed by a repeated evaluation in November 2024 to examine potential changes in translation quality over time. The initial evaluation utilized GPT-3.5 Turbo, whereas the follow-up employed GPT-4o.

A direct translation method was applied, using the prompt: "Please translate from English to Croatian" or "Please translate from Croatian to English", depending on the source and target language. For consistency, each prompt was entered in a new conversation thread to ensure that the model responded without influence from previous interactions.

2.3 Evaluation Procedure

Translation accuracy was evaluated by a professional translator at three levels: (1) translation of the idiom in isolation, (2) translation of an author-constructed excerpt, and (3) translation of an authentic excerpt containing the idiom.

For example excerpts, translation accuracy was assessed at two levels: (1) the idiom itself and (2) the entire excerpt. This distinction was necessary because an excerpt could be translated correctly overall while the idiom was mistranslated, or conversely, the idiom could be rendered accurately while the excerpt contained grammatical errors or conveyed an incorrect meaning.

3 Results

The results obtained when translating from English to Croatian are shown in Fig. 1, and those from Croatian to English in Fig. 2.

The translation of sole idioms from English to Croatian via ChatGPT in the first research (R1) was accurate 87.50% (complete eq.), 75% (partial eq.), and 45.83% (zero eq.) of times. In the second research (R2) conducted six months later, 87.50% (complete eq.), 45.83% (partial eq.), and 62.50% (zero eq.) of the idioms were translated correctly. In the R1, the idioms translated as parts of AOC excerpts reached the accuracy levels of 83.30% (complete eq.), 66.67% (partial eq.), and 54.17% (zero eq.). In the R2, on the other hand, the idioms in the same AOC excerpts were translated accurately 79.17% (complete eq.), 37.50% (partial eq.), and 50% (zero eq.) of times.

Idioms translated in the scope of R1 as parts of corpus excerpts were translated correctly in 75% (complete eq.), 70.83% (partial eq.), and 58.33% (zero eq.) of cases. The results of the translation of idioms as parts of corpus excerpts in R2 were: 75% (complete eq.), 50% (partial eq.), and 45.83% (zero eq.) of accuracy.

The percentage of correctly translated AOC excerpts containing idioms in the scope of R1 amounted to 70.83% (complete eq.), 58.33% (partial eq.), and 45.83% (zero eq.), while the percentage of correctly translated AOC excerpts containing idioms in the scope of R2 amounted to 62.50% (complete eq.), 41.67% (partial eq.), and 45.83% (zero eq.). Corpus excerpts containing idioms that were translated in R1 had the accuracy levels of 58.30% (complete eq.), 54.17% (partial eq.), and 50% (zero eq.). In R2, they were translated correctly in 66.66% (complete eq.), 41.67% (partial eq.), and 33.30% (zero eq.) of the cases.

The translation of sole idioms from Croatian to English via ChatGPT presented the following results: 95.83% (complete eq.), 91.67% (partial eq.), and 70.83% (zero eq.) were translated accurately in R1, and 91.67% (complete eq.), 66.67% (partial eq.), and 45.83% (zero eq.) were translated accurately in R2.

Moreover, when the idioms were translated as parts of AOC excerpts, the results of the R1 displayed the accuracy levels of 91.67% (complete and partial eq.), and 83.33% (zero eq.), while the accuracy levels obtained in the translation of idioms in R2 amounted to 95.83% (complete eq.), 75% (partial eq.), and 37.50% (zero eq.).

In R1, idioms as parts of corpus excerpts were translated accurately 95.83% (complete eq.), 91.67% (partial eq.), and 83.33% (zero eq.) of the times. In R2, on the other hand, the idioms as parts of corpus excerpts were translated accurately in 87.50%(complete eq.), 75% (partial eq.), and 58.33% (zero eq.) of the instances.

When it comes to the translation of the AOC excerpts containing idioms, the results of R1 demonstrated that 91.67% (complete eq.), 87.50% (partial eq.), and 79.17% (zero eq.) were translated accurately, while the results of R2 displayed that 95.83% (complete eq.), 79.16% (partial eq.), and 58.33% (zero eq.) of the AOC excerpts were translated accurately. Finally, the percentage of the accurately translated corpus excerpts containing idioms in the scope of R1 amounted to 91.67% (complete eq.), and 83.33% (partial and zero eq.), while the percentage of the accurately translated corpus excerpts containing idioms obtained in the R2 amounted to 83.30% (complete eq.), and 66.67% (partial and zero eq.).

The results of a longitudinal study demonstrate a clear drop across all categories, for both translation directions (from English to Croatian and vice-versa) (Fig. 3).

The decrease in all three categories for the translation direction from English to Croatian is as follows: from 74.99% to 74.17% for complete-equivalence idioms, from 65% to 43.33% for partial-equivalence idioms, and from 50.83% to 47.49% for zero-equivalence idioms.

The decrease in the accuracy results when translating from Croatian to English, on the other hand, is as follows: from 83.33% to 81.74% for complete-equivalence idioms, from 89.17% to 72.50% for partial-equivalence idioms, and from 80% to 53.33% for zero-equivalence idioms.

McNemar's test is used to compare the accuracy of GPT-3.5 Turbo and GPT-4o model outputs for the translations of all 72 idioms. The test statistic

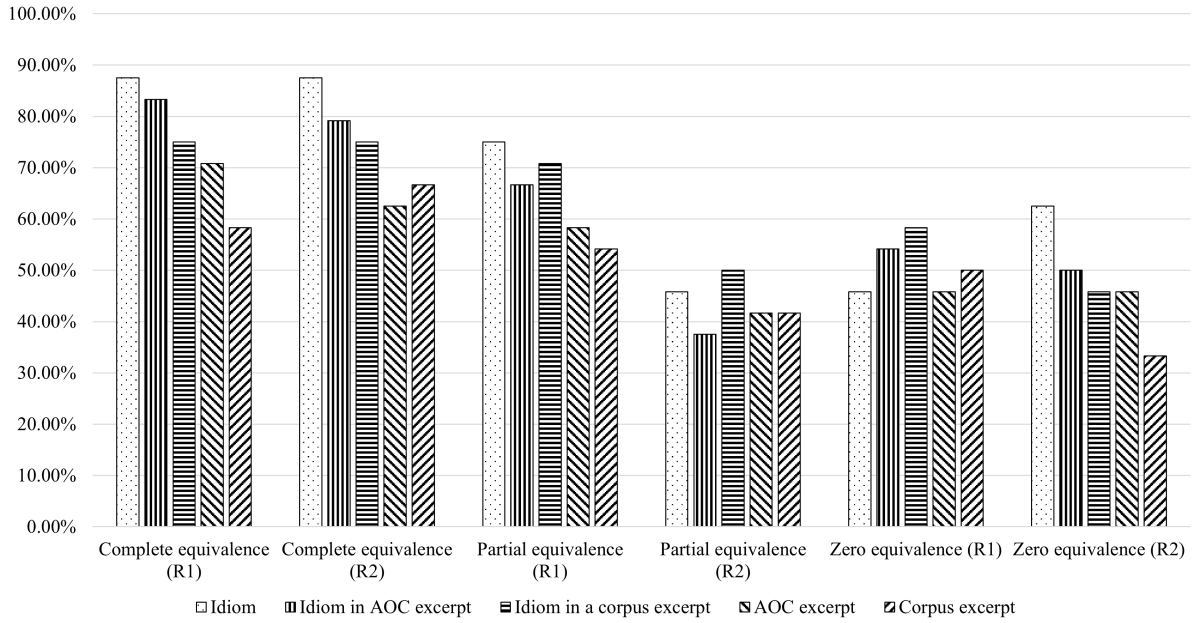

Figure 1: English-to-Croatian idiom translation accuracy.

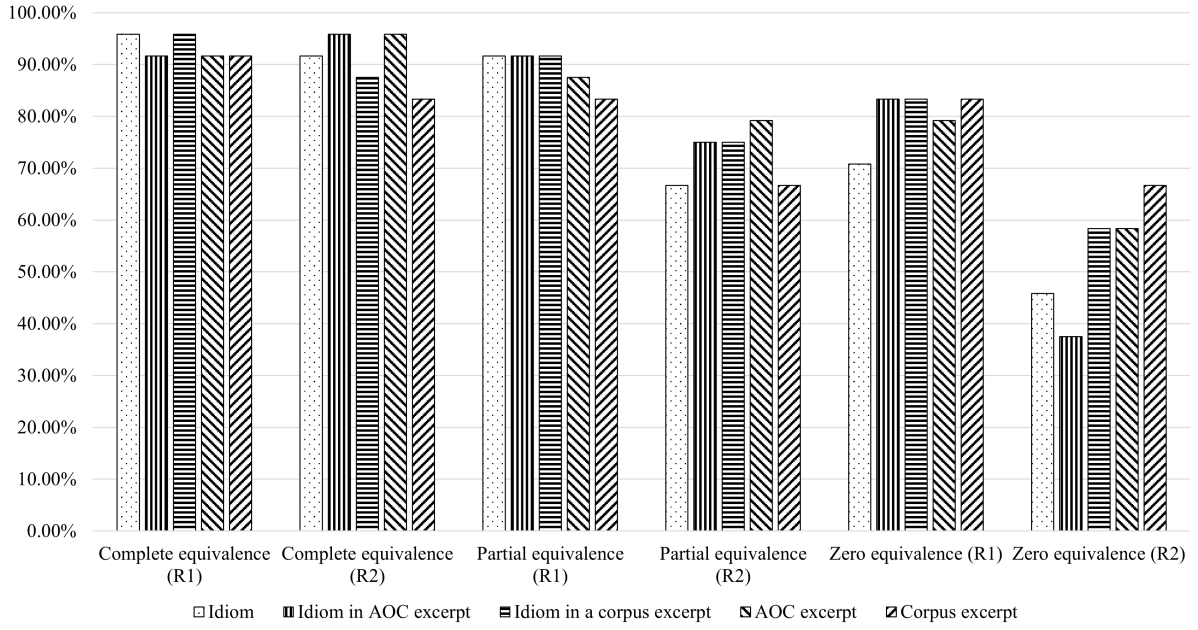

Figure 2: Croatian-to-English idiom translation accuracy.

was calculated with the continuity correction. For the translations of English idioms, the chi-square value was 5.76 with 1 degree of freedom, yielding a two-tailed p value of 0.0164. For the translations of Croatian idioms, the chi-square value was 8.47 with 1 degree of freedom, yielding a two-tailed p value of 0.0036. In both cases, the differences are statistically significant.

4 Discussion and Conclusion

As expected, the level of equivalence in idiom structure appears to play a significant role in the accuracy of idiom translation and the translation of texts containing idioms. Idioms and excerpts classified as having complete equivalence achieved the highest accuracy rates, while accuracy decreased as equivalence declined. Consequently, idioms with partial equivalence were, on average, translated less accurately than those with complete equivalence,

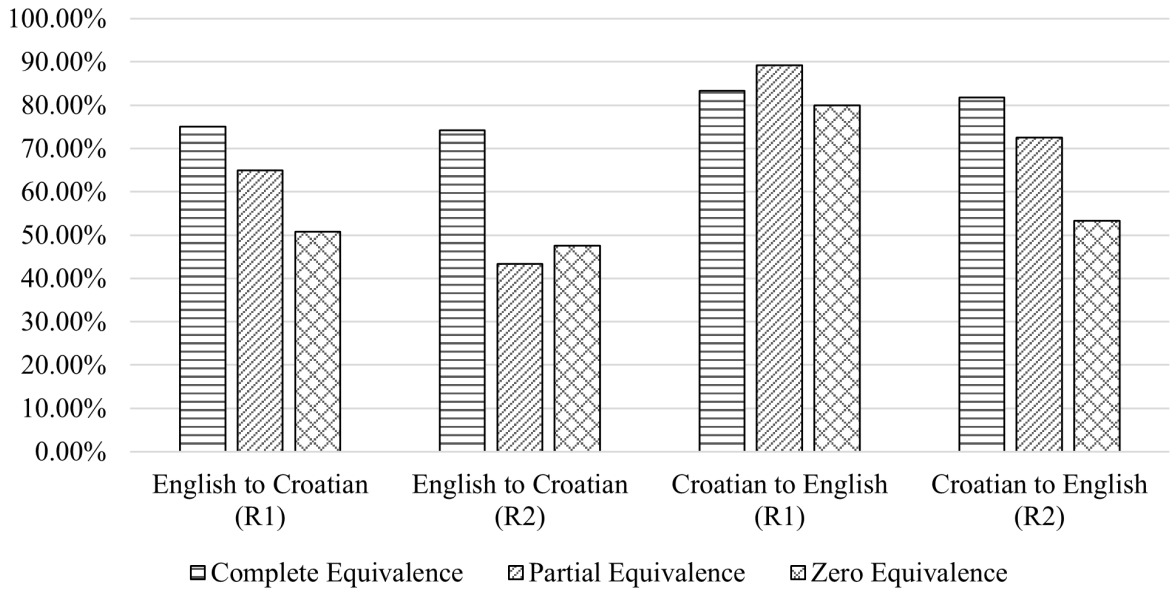

Figure 3: Longitudinal study average translation accuracy per category and translation direction.

and idioms with zero equivalence were the least accurately translated of the three categories.

Translation accuracy was also influenced by the source and target language. The findings indicate that translations from Croatian to English, i.e., from a moderately resourced language to a highly resourced language, were generally more accurate than translations in the opposite direction.

In contrast, the presence of context did not appear to have a significant impact on accuracy. Similarly, no meaningful difference was observed between translations of excerpts drawn from authentic corpus sources and those created by the author.

What follows is a concise analysis of the errors observed. Out of the four strategies for translating idioms identified by Adelnia and Dastjerdi Adelnia and Dastjerdi (2011), we observe the use of idioms equivalent in both meaning and form, idioms equivalent in meaning but not form, and paraphrasing. We did not find cases where the idiom was omitted altogether. Inaccurate translations mostly resulted from literal translations (Table 1).

Cases in which GPT-4o failed to convey the intended meaning and in which GPT-3.5 Turbo outperformed GPT-4o are presented in Table 2. GPT-3.5 Turbo avoided literal translations by providing an equivalent expression or by resorting to paraphrasing.

At times, non-literal translations also fell short of transferring the intended meaning (Table 3). In certain instances, both successfully transferred the meaning, though employing different strategies (Table 4).

Future research could expand the pool of experts involved in both the classification of idioms and the evaluation of translation accuracy. For example, a larger group of linguists could collaborate to reach consensus on categorizing idioms as complete, partial, or zero equivalence. Likewise, involving more language professionals to independently assess translation accuracy could enhance objectivity by incorporating a broader range of perspectives.

Another promising avenue would be to conduct a preliminary study in which professional translators propose their own equivalents for the idioms used in the research. These expert-provided translations could serve as a benchmark for evaluating MT output, allowing comparisons between MT-generated translations and professional standards. This approach would not only improve evaluation accuracy but also yield insights into how closely MT tools align with expert translation practices.

Finally, further longitudinal research could provide valuable data for tracking improvements over time and identifying persistent challenges associated with specific types of idioms.

5 Ethical Considerations and Limitations

This study adhered to ethical research practices in its use of AI models, data, and evaluation met-

AOC in English	GPT3.5 Turbo	GPT-4o
She had **to break the ice** at the competition with her song.	”Morala je **razbiti led** na natjecanju svojom pjesmom.”	”Morala je **razbiti led** na natjecanju svojom pjesmom.”
I **broke my neck trying** to make her happy, and all for nothing.	”Pokušavajući je usrećiti, **slomio sam si vrat**, a sve uzalud.”	”**Slomio sam vrat** pokušavajući je usrećiti, a sve uzalud.”
I am willing **to pay a king's ransom** for an apartment, as long as I have my own peace.	”Spreman sam **platiti kraljevski otkup** za stan, sve dok imam svoj mir.”	”Spreman sam **platiti kraljevski otkup** za stan, sve dok imam svoj mir.”

Table 1: Examples of literal translations.

AOC in English	GPT3.5 Turbo	GPT-4o
I **was on cloud nine** when I heard the news.	”**Bio sam na sedmom nebu** kad sam čuo vijesti.”	”**Bio sam na devetom nebu** kad sam čuo vijest.”
Her indifference to my pain **was the last straw**, I decided there and then that we could no longer be friends.	”Njena ravnodušnost prema mojoj boli **bila je kap koja je prelila čašu**, tada sam odlučio da više ne možemo biti prijatelji.”	”Njezina ravnodušnost prema mojoj boli **bila je posljednja kap**, tada i tamo sam odlučio da više ne možemo biti prijatelji.”
Working full-time while studying is a nightmare, but **no pain, no gain**.	”Raditi puno radno vrijeme dok studiraš je noćna mora, ali **nema uspjeha bez odricanja**.”	”Raditi puno radno vrijeme dok studiraš je noćna mora, ali **bez bola, nema dobitka**. ”
I consider her my best friend, even though we only see each other **once in a blue moon**.	”Smatram je svojom najboljom prijateljicom, iako **se rijetko vidimo**.”	”Smatram je svojom najboljom prijateljicom, iako se viđamo **samo jednom u plavom mjesecu**.”

Table 2: Examples where GPT3.5 Turbo outperformed GPT-4o.

AOC in Croatian	GPT3.5 Turbo	GPT-4o
”Poslije posla obično imamo **sjedeljku**, iako zapravo ne možemo dočekati polazak kući.”	After work, we usually have a **chitchat**, even though we can't wait to go home.	After work, we usually have a **meeting**, although we actually can't wait to go home.

Table 3: Example of mistranslations.

AOC in English	GPT3.5 Turbo	GPT-4o
No matter how dishonest people around him are, he always **wears his heart on his sleeve**.	”Ma koliko ljudi oko njega bili nepošteni, uvijek **otvoreno pokazuje svoje osjećaje**.”	”Bez obzira koliko nepošteni ljudi oko njega bili, on uvijek **nosi srce na dlanu**.”

Table 4: Different translation strategies.

hods. The experiments were conducted using the proprietary AI model OpenAI GPT-3.5 Turbo and GPT-4o, accessed under its official terms of service without any attempt to circumvent licensing restrictions or reverse-engineer the system. The corpus excerpts were obtained from the open web, which may include content with varying licensing conditions. To mitigate potential concerns, only short excerpts were used strictly for research purposes, and no redistribution of raw data is intended.

Despite these precautions, several limitations must be acknowledged. One of the limitations is the reliance on a single evaluator, which precluded the assessment of inter-annotator agreement. In future work, we plan to engage multiple evaluators and systematically compute inter-annotator agreement to strengthen the reliability of our findings. Final judgments will be determined by majority vote among the evaluators.

Secondly, the study utilised a relatively small dataset consisting of 24 Croatian and 24 English idioms per category and a total of 72 idioms per each translation direction. This small dataset size may affect the generalisability of the study's findings.

Lastly, an inherent limitation of working with proprietary AI systems is the lack of transparency regarding software updates and the potential influence of prompt design, both of which may affect reproducibility and comparability of results. In this study, we relied exclusively on the free tier available at the time of the assessments. Future work will include a comparison between free and paid plans to examine potential performance differences.

References

Amineh Adelnia and Hossein Vahid Dastjerdi. 2011. Translation of idioms: A hard task for the translator. *Theory and practice in language studies*, 1(7):879–883.

Mona Baker. 1992. *In other words: A coursebook on translation*. Routledge.

Christos Baziotis, Prashant Mathur, and Eva Hasler. 2022. Automatic evaluation and analysis of idioms in neural machine translation. *arXiv preprint arXiv:2210.04545*.

Sundesh Donthi, Maximilian Spencer, Om Patel, Joon Yong Doh, Eid Rodan, Kevin Zhu, and Sean O'Brien. 2025. Improving llm abilities in idiomatic translation. In *Future of Information and Communication Conference*, pages 361–375. Springer.

ELIS. 2024. European language industry survey 2024. trends, expectations and concerns of the european language industry. Technical report.

ELIS. 2025. European language industry survey 2024. trends, expectations and concerns of the european language industry. Technical report.

Rosemarie Gläser. 1984. The translation aspect of phaseological units in english and german. *Papers and Studies in Contrastive Linguistics Poznan*, 18:123–134.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2024. Translate meanings, not just words: Idiomkb's role in optimizing idiomatic translation with language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18554–18563.

Maja Manojlović, Luka Dajak, and Marija Brkić Bakarić. 2017. Idioms in state-of-the-art croatian-english and english-croatian smt systems. In *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1546–1550. IEEE.

Yafei Zhu, Daisy Monika Lal, Sofiia Denysiuk, and Ruslan Mitkov. 2024. From neural machine translation to large language models: Analysing translation quality of chinese idioms. In *Proceedings of the New Trends in Translation and Technology Conference*, pages 247–260, Shoumen, Bulgaria. INCOMA Ltd.

A Appendix

Idiom in English	Idiom in Croatian
without batting an eye	okom da ne trepnem
to have one's head in the clouds	biti glavom u oblacima
keep a cold head	sačuvati hladnu glavu
get under someone's skin	uvući se nekome pod kožu
turn one's back on someone	okrenuti nekome leđa
be worried to death	biti smrtno zabrinut
leave a bitter taste in one's mouth	ostaviti gorak okus u ustima
welcome with open arms	dočekati raširenih ruku
be fed up with	biti sit nekoga/nečega
to believe the glass is half empty	misliti da je čaša napola prazna
as clear as day	jasno kao dan
to take something with a grain of salt	uzeti što sa zrnom soli
divide and conquer	podijeli pa vladaj
to break the ice	probiti led
on thin ice	na tankom ledu
once and for all	jednom za svagda
it's the least I can do	to je najmanje što mogu učiniti
or something like that	ili tako nešto
it all comes to the same thing	sve se svodi na isto
be no better than	ne biti ništa bolji od
collect dust	skuplja prašinu
crocodile tears	krokodilske suze
doesn't hold water	ne drži vodu
don't look a gift horse in the mouth	poklonjenom konju se ne gleda u zube

Table A1: Complete equivalence.

Idiom in English	**Idiom in Croatian**
to be loaded	biti pun love, biti pun kao brod
break one's neck	pretrgati se
wear one's heart on one's sleeve	nositi srce na dlanu
from the bottom of one's heart	od sveg srca
to have one's heart in one's mouth	imati srce u petama
to quake in one's boots	tresti se od straha
to put one's foot down	lupiti šakom o stol
to make one's skin crawl	prolaze me trnci
to be head over heels in love	biti zaljubljen do ušiju
to lose one's temper	izgubiti živce
to get something out of one's system	izbaciti što iz sebe
to pull one's hair out	čupati si kosu
to be on cloud nine	biti na sedmom nebu
hold one's tongue	držati jezik za zubima
be like a bull in a china shop	biti kao slon u staklani
to be fit as a fiddle	biti zdrav kao dren
to promise the moon	obećati brda i doline
every now and then	svako toliko
it's the same old story	uvijek ista priča
a hot potato	goruća tema
to be one's flesh and blood	biti nečija krv
for goodness' sake	za boga miloga
no pain no gain	bez muke nema nauke
blood is thicker than water	krv nije voda

Table A2: Partial equivalence.

Idiom in English	**Idiom in Croatian**
a king's ransom	brdo love
a bull session	sjedeljka
shoot the bull	govoriti kao navijen
to be out to lunch	biti odsutan duhom
to lose heart	klonuti duhom
to have bats in one's belfry	imati mušice u glavi
to carry a torch for someone	imati tihu patnju
look as though butter would not melt in your mouth	praviti se nedužnim
would not say boo to a goose	bojati se vlastite sjene
to be rolling in the aisles	pucati od smijeha
to be the last straw	biti kap koja je prelila čašu
to hold one's horses	stati na loptu
it's raining cats and dogs	lijeva kao iz kabla
to pull no punches	nemati dlake na jeziku
once in a blue moon	svake prijestupne godine
if you've seen one, you've seen them all	svi su ti oni isti
and what have you	i što sve ne
when pigs fly	kad na vrbi rodi grožđe
miss the boat	prošla baka s kolačima
to cut to the chase	prijeći na stvar
to be a dead ringer	biti pljunut (netko)
different strokes for different folks	sto ljudi, sto ćudi
doesn't know beans about it	nema blage veze
to eat crow	posuti se pepelom

Table A3: Zero equivalence.