It's Not a Walk in the Park!
Challenges of Idiom Translation in Speech-to-text Systems

Iuliia Zaitova[1], Badr M. Abdullah[1], Wei Xue[2,1]
Dietrich Klakow[1], Bernd Möbius[1], Tania Avgustinova[1]
[1]Saarland University, Germany
[2]Xi'an Jiaotong University, China
izaitova@lsv.uni-saarland.de

Abstract

Idioms are defined as a group of words with a figurative meaning not deducible from their individual components. Although modern machine translation systems have made remarkable progress, translating idioms remains a major challenge, especially for speech-to-text systems, where research on this topic is notably sparse. In this paper, we systematically evaluate idiom translation as compared to conventional news translation in both text-to-text machine translation (MT) and speech-to-text translation (SLT) systems across two language pairs (German to English, Russian to English). We compare state-of-the-art end-to-end SLT systems (SeamlessM4T SLT-to-text, Whisper Large v3) with MT systems (SeamlessM4T SLT-to-text, No Language Left Behind), Large Language Models (DeepSeek, LLaMA) and cascaded alternatives. Our results reveal that SLT systems experience a pronounced performance drop on idiomatic data, often reverting to literal translations even in higher layers, whereas MT systems and Large Language Models demonstrate better handling of idioms. These findings underscore the need for idiom-specific strategies and improved internal representations in SLT architectures.

1 Introduction

> "The difference between the right word and the almost right word is really a large matter – it's the difference between lightning and a lightning bug."
>
> —Mark Twain

Imagine explaining to someone unfamiliar with English that it is *"raining cats and dogs"* or that you are feeling *"under the weather."* Although idioms carry meanings that cannot be derived from the meaning of individual words alone, humans can easily interpret them by relying on context and cultural knowledge. However, machine translation systems often produce literal, incorrect or nonsensical translations (Dankers et al., 2022; Baziotis et al., 2023; Rambelli et al., 2023; Tian et al., 2023).

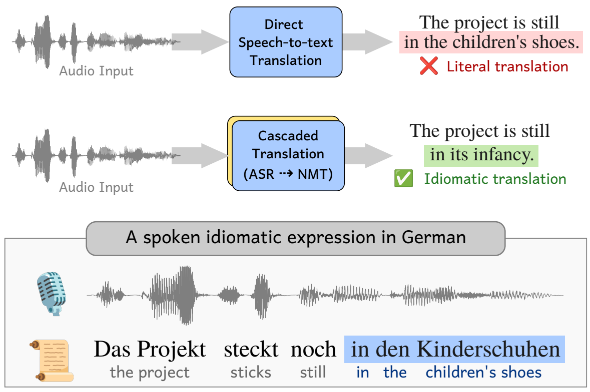

Figure 1: An illustrated example of translating a spoken idiomatic expression. The German idiom *"in den Kinderschuhen"*—literally translates to *"in children's shoes"*—means something is in its beginning stages, equivalent to the English *"in its infancy."* In this paper, we systematically assess the performance of two modes of spoken language translation for idiom translation: (1) direct speech-to-text translation, and (2) cascaded speech translation whereby the audio is first transcribed by a ASR system followed by a text-based machine translation.

Prior work has extensively examined idiom translation in text-based machine translation (MT) systems (Boisson et al., 2022; Avram et al., 2023; Liu et al., 2023; Bui and Savary, 2024), yet the topic of idioms in speech translation has received comparatively little attention. Despite the success of speech translation systems such as SeamlessM4T (Barrault et al., 2023, 2025) and Whisper (Radford et al., 2022), which achieve state-of-the-art results across many languages and acoustic conditions, speech translation systems might be particularly prone to failing on idiomatic content due to the additional complexity of integrating acoustic, syntactic, and semantic information. Understanding if and why such failures occur is essential to further improving speech-to-text translation (SLT) systems.

In this paper, we provide the first systematic com-

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 31310–31322
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

parison of idiom translation in MT, general purpose Large Language Models (LLMs), and SLT for German→English and Russian→English language pairs. We investigate:

- The relative performance of end-to-end SLT (SeamlessM4T for audio, Whisper Large v3) vs. MT (SeamlessM4T for text, No Language Left Behind), general-purpose LLMs (LLaMA 3, DeepSeek-v3), and cascaded approaches.
- How these systems handle idiomatic and news data, as measured by both COMET (Rei et al., 2020) and human annotation.
- Layer-wise performance of MT and SLT systems via DecoderLens analysis (Langedijk et al., 2024) to pinpoint how and at which encoder layers these systems fail on idioms.

Our experiments reveal that SLT significantly underperforms MT and Large Language Models (LLMs) on idiomatic data, even though they perform competitively on conventional news text. We make our code, evaluation datasets, and their annotated subsets publicly available at `https://github.com/IuliiaZaitova/idiom_s2t`.

2 Related Work

Idiom Translation in text-based systems. The difficulty of translating and handling idioms has been extensively studied in MT systems and LLMs. For instance, Dankers et al. (2022) and Baziotis et al. (2023) explored how Transformer architectures handle figurative language, identifying a tendency to produce literal translation.

Strategies such as fine-tuning on idiom-focused parallel data (Boisson et al., 2022; Avram et al., 2023) have shown promising improvements in idiom translation accuracy, though translation systems remain vulnerable to varied contexts and domains.

SLT Systems. SLT has seen significant advances with recent end-to-end architectures such as Whisper (Radford et al., 2022) and SeamlessM4T (Barrault et al., 2023, 2025). Earlier SLT research often relied on cascaded approaches, combining an automatic speech recognition (ASR) module with a separate MT system (Niehues et al., 2018; Iranzo-Sánchez et al., 2021). Recently, cascaded speech-to-text translation models have encountered criticism due to an intrinsic shortcoming of 'error propagation'. Techniques were proposed to mitigate this shortcoming and enhance the accuracy of the translation in cascaded systems (Min et al., 2025). However, the IWSLT 2023 Evaluation Campaign (Agarwal et al., 2023) still notes that cascaded approaches remain competitive in certain scenarios. These systems often outperform end-to-end systems when leveraging high-resource ASR and MT components, especially for languages with limited training data for direct SLT.

Evaluation of Figurative Language Translation. Song and Xu, 2024 explore which automatic metrics work best for evaluating multiword expressions (MWEs) and figurative language in translation. They conclude that surface-level string metrics like BLEU (Papineni et al., 2002) often fail to capture nuanced meaning shifts in idiomatic data, whereas semantic metrics like COMET (Rei et al., 2020) correlate more reliably with human judgments of MWE translation quality.

Interpretability and Layer-wise Analysis. In parallel with improvements in model performance, interpretability methods seek to reveal *how* and *where* complex systems process inputs. Voita et al. (2019) and Clark et al. (2019) examine attention heads in Transformer models, showing that syntactic and semantic information is often distributed across multiple layers. More recently, Langedijk et al. (2024) proposed DecoderLens analysis, which replaces a model's final encoder output with intermediate layer representations, translating them to human-readable text. This method offers deeper insight into how the output evolves throughout the encoding process, which is particularly useful for diagnosing issues of incorrect translation.

3 Methodology

3.1 Task and Scope

Idioms present unique challenges in translation due to their non-literal nature, which often requires contextual and cultural understanding. We focus on translating idiomatic and, for contrast, conventional news datasets in two language pairs (German→English, Russian→English) across speech and text modalities.

3.2 Systems Evaluated

MT Systems

1. SeamlessM4T (text-to-text) with version `facebook/seamless-m4t-v2-large`: A state-of-the-art multilingual MT system

capable of direct text-to-text translation across multiple languages.

2. No Language Left Behind (NLLB) with version `facebook/nllb-200-3.3B`: A system developed for enhancing translation quality in low-resource languages, capable of translating over 202 different languages with state-of-the-art results (Team et al., 2022).

Large Language Models (LLMs)

1. LLaMA 3 models fine-tuned for specific languages: (a) IlyaGusev/saiga_LLaMA3_8b (Gusev, 2025) fine-tuned for Russian, and (b) VAGOsolutions/LLaMA-3-SauerkrautLM-8b-Instruct (Solutions, 2025) fine-tuned for German.

2. DeepSeek-V3 (DeepSeek-AI et al., 2025): A multilingual LLM optimized for translation, reasoning, and code generation tasks. It tops the leaderboard among open-source models.

LLM Prompts To ensure transparency, we include the prompts used to produce translation to English by LLaMA and DeepSeek models in Appendix A.

SLT Systems

1. SeamlessM4T (speech-to-text) with version `facebook/seamless-m4t-v2-large`: An end-to-end multilingual system capable of translating speech inputs into text.

2. Whisper Large v3 with version `openai/whisper-large-v3` (Whisper): A highly robust speech recognition and translation model with 1.55 billion parameters, designed to handle diverse languages and acoustic conditions.

Cascaded Systems We formed cascaded systems by feeding audio inputs (16kHz mono WAV) into either SeamlessM4T or Whisper for ASR, then passing their transcriptions into each MT system and LLM. The transcribed text's capitalization and punctuation was retained.

3.3 Evaluation Datasets

3.3.1 Conventional News Corpus

To evaluate general translation performance, we used the professionally translated *News Commentary* parallel corpus[1]. This dataset includes formal, well-structured news text in political and economic domain with minimal use of figurative language, making it ideal as a baseline for general translation performance. By providing consistent and straightforward content, the *News Commentary* corpus allows us to contrast the performance of translation systems under conventional conditions with their ability to handle idiomatic data. To perform our evaluation, we randomly selected 250 sentences from the News Commentary corpus for both language pairs. Examples from the *News Commentary* corpus are shown below:

> **Russian:** *Что же может оправдать очередной значительный рост цен на золото, начиная с сегодняшнего дня?* (Eng. trans.: So what could justify another huge increase in gold prices from here?)

> **German:** *Damals lag Gold bei 850 Dollar, also in heutigem Geldwert um einiges über 2.000 Dollar.* (Eng. trans.: Back then, gold hit $850, or well over $2,000 in today's dollars.)

3.3.2 Idiomatic Corpus

Idiomatic data used for evaluation is sourced from the *Idioms-InContext-MT* dataset (Stap et al., 2024)[2]. From the 1,000 examples available in the dataset per language pair, we manually selected 250 idioms that require non-literal translation to preserve their figurative meaning. For instance:

> **German:** ***Es ist mir wurst**, wenn du nicht kommst.* (literally: *It is sausage to me if you don't come.*, meaning: *Eng. trans.: I couldn't care less if you don't come.*)

> **Russian:** — Ну да! Мы с тобой — **два сапога пара**! — охотно согласился Шурик. (literally: *Well, yes! You and me are like two boots of a pair!, Shurik happily agreed.*, meaning: *Eng. trans.: Well, yes! You and me are like two peas in a pod!, Shurik happily agreed.*)

We excluded idioms whose figurative meaning is preserved in a literal translation. For example:

> **Russian:** *Они и **мухи не обидят.*** (literally: *They wouldn't hurt a fly.*)

[1] https://metatext.io/datasets/news-commentary-parallel-corpus

[2] https://github.com/amazon-science/idioms-incontext-mt

Category	Description	Example (De)
Correct	*Idiomatic* †: Preserves figurative meaning	*Es ist mir wurst* → *I couldn't care less*
	Paraphrase †: Literal conversion with meaning	*Es ist mir wurst* → *It doesn't matter*
Partially Correct	Core meaning with minor errors; more than 50% of the sentence is translated correctly	*Es ist mir wurst* → *It matters to me*
Literal Translation †	Word-for-word idiom translation that loses the idiomatic meaning; the sentence translation otherwise correct	*Es ist mir wurst* → *It is sausage*
Incorrect (Relevant)	Addresses the same topic but misrepresents critical information; less than 50% of the sentence is translated correctly	*Es ist mir wurst* → *I want to go*
Incorrect (Hallucination)	Fabricated unrelated content	*Es ist mir wurst* → *I'm not a child*
Empty/Ellipsis	Missing/empty output	*Es ist mir wurst* → *„„„*

Table 1: Annotation scheme for manual translation evaluation. † marks categories specific to idiom evaluation. The German phrase *'Es ist mir wurst'* is correctly translated to English as *'I couldn't care less'*.

> **German:** *Als er die Nachricht hörte,* ***brach es ihm das Herz.*** (literally: *When he heard the news, it broke his heart.*)

This selection process ensures the focus remains on idioms that challenge machine translation systems, allowing us to evaluate their ability to translate idiom figuratively.

To enable SLT evaluation, we synthesized audio for all text segments using Microsoft Edge voice services[3], which employs neural text-to-speech (TTS) architectures comparable to state-of-the-art systems. Synthesizing speech for text-based datasets is a widely used practice in translation research (Jia et al., 2019; Moslem, 2024; Bamfo Odoom et al., 2024).

While synthetic speech may have minor deviations in prosody or emphasis (Wester et al., 2016; Chan and Kuang, 2024), such factors are secondary in idiomaticity-centered MT evaluation. Modern TTS tools have been shown to approximate natural speech quality so closely that distinguishing synthetic from human speech is non-trivial (Jiang, 2024; Ji et al., 2024). To ensure that translation differences come from the MT systems rather than acoustic variations, we used consistent female voice presets across all synthesized audio. This approach reduces variability and is in line with previous works demonstrating that consistent speaker characteristics improve the reliability of MT system evaluation (Fuckner et al., 2023).

[3] `https://www.microsoft.com/edge`

3.4 Evaluation Procedure

To assess model performance, we employed both automatic and manual evaluation methods.

3.4.1 Automatic Metrics

For the automatic evaluation of translation quality, we utilize the COMET metric (Rei et al., 2020) of version `Unbabel/wmt22-comet-da`. COMET is a state-of-the-art framework that has shown a high correlation with human judgments. It assesses translations based on semantic equivalence and fluency. This is particularly critical for idioms where literal translation fails to convey semantic equivalence, and contextual understanding is essential (Song and Xu, 2024). By using COMET, we were able to ensure that both the intended meaning and the naturalness of idioms rather than form similarity are prioritized.

3.4.2 Human Annotation of Translation Output

To supplement COMET evaluations, two human annotators evaluated a random sample of 50 translations from each language-dataset-model combination using the annotation scheme in Table 1, where categories range from *Correct* to *Empty/Ellipsis*.

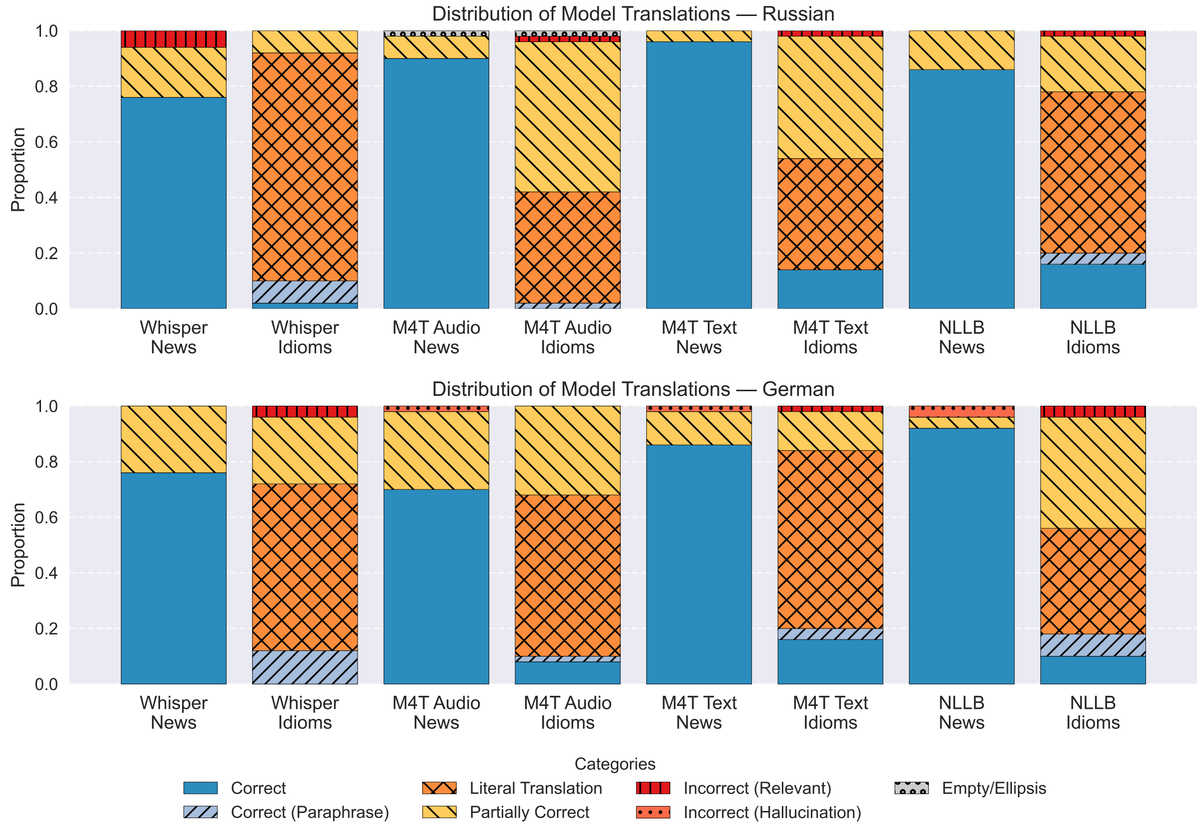

Figure 2: Distribution of translation output categories across models for German→English and Russian→English translation. Each bar represents a model's output distribution on either news or idiomatic test sets. Speech-to-text translation systems mostly show lower proportions of correct translations for idioms compared to text-to-text translation systems, indicating a particular challenge of idiom translation in speech-to-text systems.

For clear comparison, only encoder-decoder models were used for this evaluation. For idioms, annotators explicitly judged if figurative meaning was maintained by annotating correct translations as either *Correct (Idiomatic)* or *Correct (Paraphrase)*. The category *Literal Translation* was also only used in idiom translation evaluation. The annotators resolved any disagreements through discussion to ensure consistent evaluation criteria.

4 Results and Discussion

4.1 Overall Performance

Table 2 presents COMET scores for German and Russian, comparing model performance on news vs. idiom datasets. For each model, we further evaluated the differences in performance on two datasets using the Mann–Whitney U test. After applying Bonferroni correction for multiple comparisons, all models demonstrated statistically significant differences in performance on news vs. idioms with corrected p-values below 0.001 for both language pairs. Additional statistical analyses, i.e. Kruskal-Wallis tests, standard deviation, and median performance comparisons, are provided in Appendix B. The word error rates of ASR transcription used in the cascaded systems are provided in Appendix C.

MT and LLM vs. SLT: The DeepSeek model largely outperforms all other models, especially on idiom translation. Other text-based systems (including NLLB, SeamlessM4T, and LLaMA variants) consistently outperform SLT systems (SeamlessM4T and Whisper) on idiom dataset regardless of language, and only in some cases on news dataset, such as NLLB and M4T with higher COMET scores for both German and Russian.

Performance Drop on Idioms: SLT systems' COMET scores decline sharply when moving from news to idioms (e.g., a 24.2% drop from 0.844 to 0.640 in German→English for Whisper).

Cascaded Systems: Although cascaded systems do not reach the end-to-end text-based systems' performance level, they mostly outperform end-to-end SLT systems. This seems to suggest that SLT systems errors are not solely due to ASR transcription

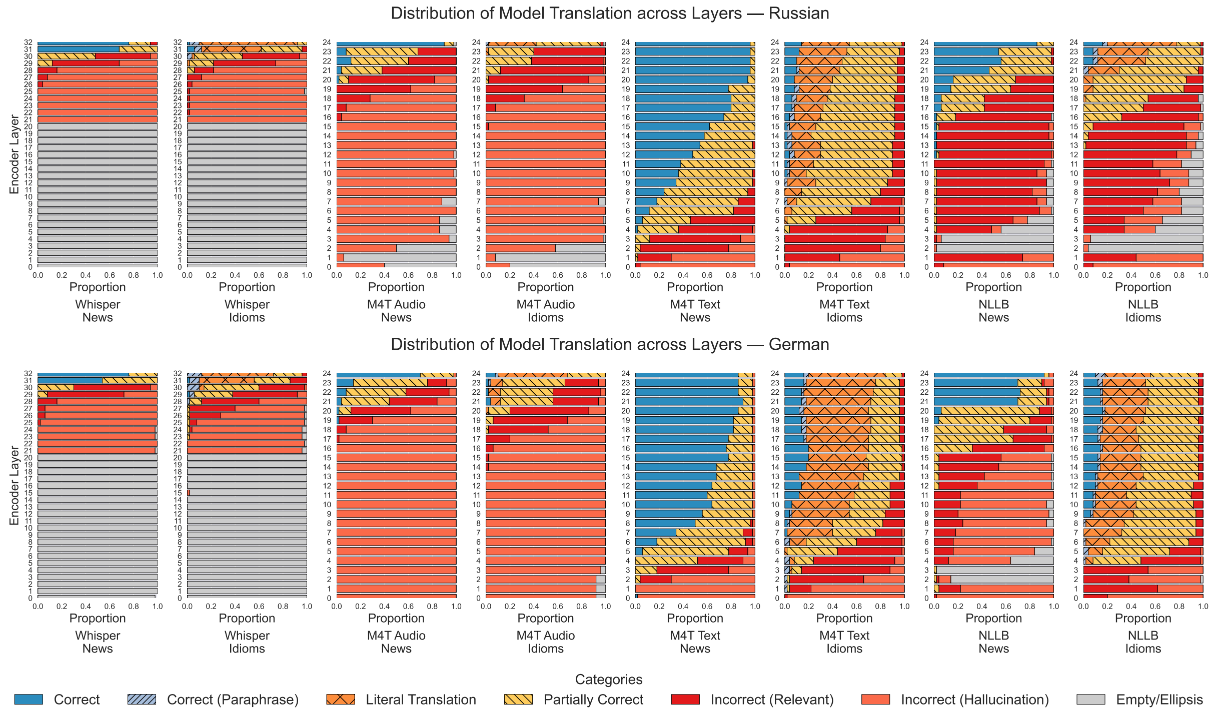

Figure 3: Distribution of translation error categories across encoder layers for German→English and Russian→English translation. Each subplot shows the evolution of translation quality through different encoder layers for a specific model and domain (news vs. idioms). The x-axis shows the proportion of translations falling into each category, and y-axis represents encoder layers.

but also reflect deeper challenges in the end-to-end systems. Such challenges may involve the integration of acoustic and semantic information, which is particularly important for semantically complex idiomatic language.

4.2 Translation Error Distributions

Figure 2 displays the distribution of translation error categories (listed in Table 1) for each encoder-decoder model in the Russian→English (top panel) and German→English (bottom panel) translations. Two SLT systems (Whisper and SeamlessM4T) and two MT systems (NLLB and SeamlessM4T) were analyzed. As shown in Figure 2, there is a clear difference in performance on news and idiom datasets. For news, both SLT and MT systems produce predominantly correct outputs. By contrast, idiomatic datasets see less correct and more divergent outputs. SLT and MT systems both produce a high proportion of the *Literal Translation* category for idiom translation. This points to a shared challenge of recognizing idioms, although it is less pronounced in MT systems. Additionally, SLT systems are more likely to generate not only literal but also partially correct translations, while MT systems demonstrate a better, though far from perfect, handling of figurative language. These results emphasize the general shortfall of translation systems in capturing idiomatic meaning. These patterns emphasize the broader challenge that idiomatic expressions pose for current translation systems, revealing fundamental limitations in their ability to capture non-literal meaning.

5 Layer-wise Analysis with DecoderLens

To understand where translation systems capture or lose idiomatic meaning, we analyzed four encoder-decoder translation systems using DecoderLens (Langedijk et al., 2024): two SLT systems (Whisper and SeamlessM4T) and two MT systems (NLLB and SeamlessM4T).

DecoderLens enables analysis of intermediate representations by replacing the final encoder output with activations from each encoder layer, allowing the decoder to attend to these intermediate states. It reveals how semantic meaning evolves through the network by converting hidden representations into human-readable text. For each model, we extracted outputs from all encoder layers and generated translations of 50 examples, which then were annotated by two human annotators using the scheme in Table 1. The results highlight key differ-

system	German → English		Russian → English	
	news	idioms	*news*	idioms
Whisper Audio Encoder				
Whisper (Direct SLT)	*0.8437*	0.6402	*0.8318*	0.6916
Whisper (ASR) → NLLB	*0.8767*	0.6774	*0.8523*	0.7180
Whisper (ASR) → Seamless (MT)	*0.8805*	0.6703	*0.8603*	0.7147
Whisper (ASR) → LLaMA	*0.8685*	0.6875	*0.8438*	0.7339
Whisper (ASR) → DeepSeek	*0.8887*	**0.7584**	*0.8607*	**0.7873**
Seamless M4T Audio Encoder				
Seamless (Direct SLT)	*0.8697*	0.6483	*0.8512*	0.6941
Seamless (ASR) → NLLB	*0.8672*	0.6790	*0.8594*	0.7025
Seamless (ASR) → Seamless (MT)	*0.8729*	0.6719	*0.8614*	0.7185
Seamless (ASR) → LLaMA	*0.8624*	0.6871	*0.8454*	0.7283
Seamless (ASR) → DeepSeek	*0.8857*	**0.7635**	*0.8667*	**0.7804**
Text MT (upper bound performance)				
Seamless (Text MT and LLM)	*0.8870*	0.6784	*0.8694*	0.7262
NLLB	*0.8841*	0.6749	*0.8664*	0.7214
LLaMA	*0.8724*	0.6971	*0.8211*	0.7354
DeepSeek	*0.8940*	**0.7675**	*0.8741*	**0.7939**

Table 2: Performance comparison of translation systems across modalities and approaches, showing COMET scores for both news and idiomatic content in German→English and Russian→English translation.

ences between SLT and MT systems in processing figurative language.

5.1 Results of Layer-wise Analysis with DecoderLens

Table 3 presents an example of layer-by-layer English translation outputs from Whisper SLT system via DecoderLens for a Russian idiomatic item. As shown in the example, Layers 0–20 consistently produce empty or punctuation-only strings, indicating that the model has yet to form a meaningful textual output. Starting from Layer 21, the system attempts to generate text but mostly produces *Hallucinations*. Only in the last few layers does the system start to align with the original text (reflected by the *Incorrect but Relevant* category), and eventually produce a *Partially Correct* output at Layer 31. However, Layer 32 only manages to output a *Literal Translation*, further showing that the model fails to preserve the figurative sense of the idiom 'still waters.'

Figure 3 shows a layer-by-layer breakdown of translation outputs for 50 examples of each evaluated system using the DecoderLens method, based on the categories introduced in Table 1. Each subplot corresponds to a particular system and domain (news vs. idioms) for Russian and German.

For all the available data, direct SLT systems start to produce meaningful translations only in higher encoder layers. From there, they gradually improve from producing *Partially Correct* outputs to *Paraphrased*, *Literal*, and *Correct* translations in the final layers. For relatively straightforward news text, the model progressively refines its representations towards correct translations. By contrast, when translating idioms, SLT systems are more prone to literal translations, with only minor improvements in higher layers.

MT systems also have difficulties moving away from literal translations for idiomatic inputs. In general, however, their transitions across layers are smoother, which indicates a different internal strategy for capturing semantics.

5.2 Cross-Language Differences

Although both German and Russian see drops in idiomatic performance, German has a larger gap (0.198 on average) between news and idioms, while Russian's gap is around 0.143. The ranking of systems, however, is mostly consistent across the two languages.

Layer(s)	Output (DecoderLens)	Annotation
0–1	*(empty)*	Empty/Ellipsis
2–10	*(empty)*	Empty/Ellipsis
11	,,	Empty/Ellipsis
12–20	*(empty)*	Empty/Ellipsis
21	The world is full of people.	Incorrect (Hallucination)
22	The first part of the video is about the first part of the video.	Incorrect (Hallucination)
23	The next day.	Incorrect (Hallucination)
24	The story of the story of the story of the story of the story of . . . (truncated)	Incorrect (Hallucination)
25	The story of the man who was killed by a woman.	Incorrect (Hallucination)
26	The T-shirt of the Russian national team.	Incorrect (Hallucination)
27–28	The End	Incorrect (Hallucination)
29	You know, I'm a man of my word.	Incorrect (Relevant)
30	In the quiet you know who is the best. In the quiet you know who is the best.	Incorrect (Relevant)
31	In the silent moment you know who is in charge. The appearance and manner of a person can be deceiving.	Partially Correct
32	In a quiet room you know who is leading. The appearance and manner of human behavior can be deceiving.	Literal Translation

Table 3: **DecoderLens Layer-by-layer Outputs: Whisper for a Russian Idiomatic Sentence.**
Original (Russian): "В тихом омуте сам знаешь кто водится. . . Внешность и манера поведения человека бывают обманчивы."
Gold translation (English): "You know what they say about still waters. A person's appearance and behavior can be deceptive."

6 Conclusion

In this work, we provide a systematic comparison of speech-to-text (SLT), text-to-text machine translation (MT), and Large Language Models (LLMs) when challenged with idiomatic datasets in German→English and Russian→English. Our findings reveal the following:

1. **SLT underperforms for idioms.** Both SLT and MT systems struggle with idiomatic translation, as reflected by performance drops of COMET scores on idiom dataset compared to conventional news. Notably, the performance gap between news and idiom datasets is more pronounced for SLT systems, while MT and LLMs are better at translating idioms.

2. **Layer-wise analysis highlights structural differences between speech-to-text and text-to-text systems.** Using DecoderLens, we observe that SLT systems only start to 'refine' their translation output in the higher layers, and revert to literal translation more frequently even in higher encoder layers. MT systems, on the other hand, show a gradual improvement in capturing the intended sense when moving from intermediate to higher layers.

Overall, our study shows that translating idioms remains a bigger challenge for SLT systems compared to MT, LLMs, and cascaded systems. Although SeamlessM4T and Whisper perform competitively on conventional news, cascaded approaches combining strong ASR and text-based components provide better handling of figurative language, likely due to text-based systems' stronger semantic processing. These findings highlight the need for idiom-specific strategies and improved representations of idioms in SLT systems. For practical applications, we recommend using cascaded systems when translating speech likely to contain idiomatic expressions. We hope this study will inspire further research on figurative language in speech translation.

Ethical statement

The annotators involved in this study were compensated for their work on hourly basis.

Limitations

Annotating translation output is inherently subjective. Morever, our approach focuses only on translating German and Russian to English, while idiom usage varies widely across languages. The use of synthetic speech may differ from real-world spontaneous speech though prior work suggests minimal impact on core translation errors. Finally, DecoderLens analysis is limited to encoder-decoder architectures and may not capture idiom handling in purely decoder-based systems like LLaMA.

Acknowledgments

We thank Mutahar Aamir for his help with the data annotation, as well as the anonymous reviewers for their constructive feedback. This research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project-ID 232722074 – SFB 1102.

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutai Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Kr. Ojha, John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Andrei Avram, Verginica Barbu Mititelu, and Dumitru-Clementin Cercel. 2023. Romanian multiword expression detection using multilingual adversarial training and lateral inhibition. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 7–13, Dubrovnik, Croatia. Association for Computational Linguistics.

Bismarck Bamfo Odoom, Nathaniel Robinson, Elijah Rippeth, Luis Tavarez-Arce, Kenton Murray, Matthew Wiesner, Paul McNamee, Philipp Koehn, and Kevin Duh. 2024. Can synthetic speech improve end-to-end conversational speech translation? In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 167–177, Chicago, USA. Association for Machine Translation in the Americas.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023. Seamless: Multilingual expressive and streaming speech translation. *Preprint*, arXiv:2312.05187.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Çelebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and SEAMLESS Communication Team. 2025. Joint speech and text machine translation for up to 100 languages. *Nature*, 637(8046):587–593.

Christos Baziotis, Prashant Mathur, and Eva Hasler. 2023. Automatic evaluation and analysis of idioms in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3682–3700, Dubrovnik, Croatia. Association for Computational Linguistics.

Joanne Boisson, Jose Camacho-Collados, and Luis Espinosa-Anke. 2022. CardiffNLP-metaphor at SemEval-2022 task 2: Targeted fine-tuning of transformer-based language models for idiomaticity detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 169–177, Seattle, United States. Association for Computational Linguistics.

Van-Tuan Bui and Agata Savary. 2024. Cross-type French multiword expression identification with pretrained masked language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4198–4204, Torino, Italia. ELRA and ICCL.

Cedric Chan and Jianjing Kuang. 2024. Exploring the accuracy of prosodic encodings in state-of-the-art text-to-speech models. *Speech Prosody 2024*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT‘s attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Marcio Fuckner, Sophie Horsman, Pascal Wiggers, and Iskaj Janssen. 2023. Uncovering bias in asr systems: Evaluating wav2vec2 and whisper for dutch speakers. In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 146–151.

Ilya Gusev. 2025. Saiga llama3 8b model. `https://huggingface.co/IlyaGusev/saiga_llama3_8b`. Accessed: 2025-01-27.

Javier Iranzo-Sánchez, Javier Jorge, Pau Baquero-Arnal, Joan Albert Silvestre-Cerdà, Adrià Giménez, Jorge Civera, Albert Sanchis, and Alfons Juan. 2021. Streaming cascade-based speech translation leveraged by a direct segmentation model. *Neural Networks*, 142:303–315.

Zhoulin Ji, Chenhao Lin, Hang Wang, and Chao Shen. 2024. Speech-forensics: Towards comprehensive synthetic speech dataset establishment and analysis. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 413–421. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J. Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. 2019. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. *Preprint*, arXiv:1811.02050.

Xiaotong Jiang. 2024. Speech synthesis and quality evaluation.

Anna Langedijk, Hosein Mohebbi, Gabriele Sarti, Willem Zuidema, and Jaap Jumelet. 2024. Decoderlens: Layerwise interpretation of encoder-decoder transformers. *Preprint*, arXiv:2310.03686.

Emmy Liu, Aditi Chaudhary, and Graham Neubig. 2023. Crossing the threshold: Idiomatic machine translation through retrieval augmentation and loss weighting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15095–15111, Singapore. Association for Computational Linguistics.

Anna Min, Chenxu Hu, Yi Ren, and Hang Zhao. 2025. When end-to-end is overkill: Rethinking cascaded speech-to-text translation. *Preprint*, arXiv:2502.00377.

Yasmin Moslem. 2024. Leveraging synthetic audio data for end-to-end low-resource speech translation. *Preprint*, arXiv:2406.17363.

Jan Niehues, Ngoc-Quan Pham, Thanh-Le Ha, Matthias Sperber, and Alex Waibel. 2018. Low-latency neural speech translation. *Preprint*, arXiv:1808.00491.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Giulia Rambelli, Emmanuele Chersoni, Marco S. G. Senaldi, Philippe Blache, and Alessandro Lenci. 2023. Are frequent phrases directly retrieved like idioms? an investigation with self-paced reading and language models. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 87–98, Dubrovnik, Croatia. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

VAGO Solutions. 2025. Llama-3 sauerkrautlm 8b instruct model. `https://huggingface.co/VAGOsolutions/Llama-3-SauerkrautLM-8b-Instruct`. Accessed: 2025-01-27.

Huacheng Song and Hongzhi Xu. 2024. Benchmarking the performance of machine translation evaluation metrics with Chinese multiword expressions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2204–2216, Torino, Italia. ELRA and ICCL.

David Stap, Eva Hasler, Bill Byrne, Christof Monz, and Ke Tran. 2024. The fine-tuning paradox: Boosting translation quality without sacrificing llm abilities. *Preprint*, arXiv:2405.20089.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Ye Tian, Isobel James, and Hye Son. 2023. How are idioms processed inside transformer language models? In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 174–179, Toronto, Canada. Association for Computational Linguistics.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

M. Wester, O. Watts, and G. E. Henter. 2016. Evaluating comprehension of natural and synthetic conversational speech. In *Proceedings of Speech Prosody 2016*, pages 766–770.

A Appendix A

A.1 Prompt for LLaMA 3 fine-tuned for Russian

> You are a professional translator who translates from Russian to English.
> Only generate the target sentence, and nothing else. Follow the example below:
>
> Input sentence: У меня нет воды.
> Translation: I don't have water.
>
> Translate this sentence:

A.2 Prompt for LLaMA 3 fine-tuned for German

> You are a professional translator who translates from German to English.
> Only generate the target sentence, and nothing else. Follow the example below:
>
> Input sentence: Ich habe kein Wasser.
> Translation: I don't have water.
>
> Translate this sentence:

B Appendix B

Table 4: Performance analysis of translation models using COMET scores for German→English data

(a) **German News**: COMET score analysis for German→English translation on news data

Model	Mean	Median	Std
DeepSeek	**0.894**	0.901	0.054
Whisper + DeepSeek	0.889	0.896	0.055
M4T Text	0.887	0.894	0.059
M4T ASR + DeepSeek	0.886	0.892	0.055
NLLB	0.884	0.898	0.078
Whisper + M4T	0.880	0.889	0.062
Whisper + NLLB	0.877	0.894	0.083
M4T ASR + MT	0.873	0.883	0.066
LLaMA	0.872	0.885	0.062
M4T Audio	0.870	0.879	0.065
Whisper + LLaMA	0.869	0.882	0.067
M4T ASR + NLLB	0.867	0.884	0.084
M4T ASR + LLaMA	0.862	0.873	0.066
Whisper	0.844	0.854	0.074

Statistical Analysis:
Kruskal-Wallis H = 179.09
p-value < 2.60×10^{-31}

(b) **German Idioms**: COMET score analysis for German→English translation on idiomatic data

Model	Mean	Median	Std
DeepSeek	**0.767**	0.779	0.128
M4T ASR + DeepSeek	0.764	0.759	0.131
Whisper + DeepSeek	0.758	0.758	0.133
LLaMA	0.697	0.698	0.136
Whisper + LLaMA	0.687	0.690	0.134
M4T ASR + LLaMA	0.687	0.692	0.138
M4T ASR + NLLB	0.679	0.682	0.132
M4T Text	0.678	0.684	0.131
Whisper + NLLB	0.677	0.684	0.130
NLLB	0.675	0.665	0.130
M4T ASR + MT	0.672	0.670	0.133
Whisper + M4T	0.670	0.676	0.132
M4T Audio	0.648	0.644	0.125
Whisper	0.640	0.639	0.124

Statistical Analysis:
Kruskal-Wallis H = 275.74
p-value < 2.82×10^{-51}

Note: Models are sorted by mean COMET score. The Kruskal-Wallis test indicates statistically significant differences between model performances. The best-performing models (DeepSeek) is shown in bold.

Table 5: Performance analysis of translation models using COMET scores for Russian→English data

(a) **Russian News**: COMET score analysis for Russian→English translation on news data

Model	Mean	Median	Std
DeepSeek	**0.874**	0.878	0.051
M4T Text	0.869	0.874	0.054
M4T ASR + DeepSeek	0.867	0.871	0.054
NLLB	0.866	0.873	0.056
M4T ASR + MT	0.861	0.866	0.059
Whisper + DeepSeek	0.861	0.872	0.078
Whisper + M4T	0.860	0.868	0.063
Whisper + NLLB	0.859	0.868	0.064
M4T ASR + NLLB	0.852	0.864	0.068
M4T Audio	0.851	0.858	0.060
M4T ASR + LLaMA	0.845	0.851	0.061
Whisper + LLaMA	0.844	0.851	0.067
Whisper	0.832	0.836	0.070
LLaMA	0.821	0.858	0.122

Statistical Analysis:
Kruskal-Wallis H = 127.89
p-value < 5.49×10^{-21}

(b) **Russian Idioms**: COMET score analysis for Russian→English translation on idiomatic data

Model	Mean	Median	Std
DeepSeek	**0.794**	0.801	0.084
Whisper + DeepSeek	0.787	0.794	0.090
M4T ASR + DeepSeek	0.780	0.791	0.093
LLaMA	0.735	0.741	0.105
Whisper + LLaMA	0.734	0.737	0.103
M4T ASR + LLaMA	0.728	0.734	0.108
M4T Text	0.726	0.734	0.108
NLLB	0.721	0.736	0.117
M4T ASR + MT	0.718	0.726	0.111
Whisper + NLLB	0.718	0.735	0.115
Whisper + M4T	0.715	0.719	0.110
M4T ASR + NLLB	0.703	0.710	0.118
M4T Audio	0.694	0.699	0.116
Whisper	0.692	0.690	L0.106

Statistical Analysis:
Kruskal-Wallis H = 276.88
p-value < 1.62×10^{-51}

Note: Models are sorted by mean COMET score. The Kruskal-Wallis test indicates statistically significant differences between model performances. The best-performing model (DeepSeek) is shown in bold.

C Appendix C

System	**German**		**Russian**	
	News	Idioms	News	Idioms
Whisper	0.040	0.024	0.081	0.084
Seamless	0.085	0.037	0.124	0.145

Table 6: Average Word Error Rate of automatic speech recognition used in cascaded translation systems. Lower is better.