Miðeind at WMT25 General Machine Translation Task and Terminology Translation Task

Svanhvít Lilja Ingólfsdóttir, Haukur Páll Jónsson, Kári Steinn Aðalsteinsson, Róbert Fjölnir Birkisson, Sveinbjörn Þórðarson, Þorvaldur Páll Helgason
Miðeind ehf., Reykjavík, Iceland
mideind@mideind.is

Abstract

We present Miðeind's system contribution to two shared tasks at WMT25 – Tenth Conference on Machine Translation: The General Machine Translation Task and the WMT25 Terminology Translation Task. Erlendur is a multilingual LLM-based translation system that employs a multi-stage pipeline approach, with enhancements especially for translations from English to Icelandic. We address translation quality and grammatical accuracy challenges in current LLMs through a hybrid prompt-based approach that can benefit lower-resource language pairs. In a preparatory step, the LLM analyzes the source text and extracts key terms for lookup in an English-Icelandic dictionary. The findings of the analysis and the retrieved dictionary results are then incorporated into the translation prompt. When provided with a custom glossary, the system identifies relevant terms from the glossary and incorporates them into the translation, to ensure consistency in terminology. For longer inputs, the system maintains translation consistency by providing contextual information from preceding text chunks. Lastly, Icelandic target texts are passed through our custom-developed seq2seq language correction model (Ingólfsdóttir et al., 2023), where grammatical errors are corrected. Using this hybrid method, Erlendur delivers high-quality translations, without fine-tuning. Erlendur ranked 3rd-4th overall in the General Machine Translation Task for English-Icelandic translations, achieving the highest rank amongst all systems submitted by WMT25 participants (Kocmi et al., 2025a). Notably, in the WMT25 Terminology Shared Task, Erlendur placed 3rd in Track 1 and took first place in the more demanding Track 2 (Semenov et al., 2025).

1 Introduction

While large language models (LLMs) exhibit strong cross-lingual understanding and can produce high-quality translations from lower-resource languages into major languages like English, gaps in the models' vocabulary and limitations in their grammatical knowledge (Arnett and Bergen, 2025) often become apparent when translating into lower-resource languages (Robinson et al., 2023). Here we describe Erlendur, a multilingual LLM-based translation system designed to address these challenges by enhancing the quality and grammaticality of Icelandic translations. Our main contribution is a hybrid, multi-stage pipeline that combines preparatory text analysis, dictionary lookup, glossary integration, careful prompting, seamless handling of longer texts, and grammatical error correction.

We deployed Erlendur for our submissions to WMT25. In the General Machine Translation Task (unconstrained track)[1] for English-Icelandic translations, Erlendur achieved the highest performance among participating systems for the language pair, ranking 3rd overall behind a human translation (1st) and Gemini 2.5 Pro (2nd). Erlendur marginally outperformed GPT-4.1, though within the margin of statistical significance (Kocmi et al., 2025a). In the WMT25 Terminology Translation Task[2], where participating systems must correctly incorporate glossary terms into their translations, Erlendur placed third in Track 1, which tests injection of glossary terms into short text chunks, and secured first place in Track 2, which tests the scalability of the terminology approach, with much longer glossaries and corpus-level texts (Semenov et al., 2025).

2 System description

Erlendur (see Figure 1) is a translation service provided through Málstaður[3] (Miðeind, 2025), an integrated platform for language technology solutions aimed at Icelandic. Users can access translation

[1] https://www2.statmt.org/wmt25/translation-task.html
[2] https://www2.statmt.org/wmt25/terminology.html
[3] https://malstadur.is

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 130–135
November 8–9, 2025.

capabilities through Málstaður's editor interface, or through a speech recognition system where they can speak directly and receive translations of the transcribed text. The translation service is also provided commercially through an API.

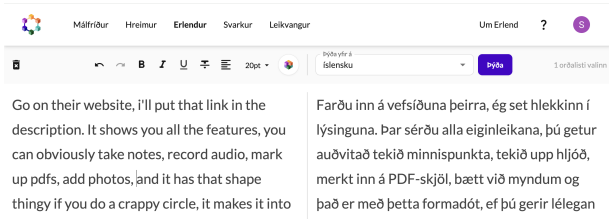

Figure 1: The Erlendur translation interface (in Icelandic) in the Málstaður platform. The user selects the target language; the source language is inferred. "1 orðalisti valinn" indicates that one glossary is selected for use during translation.

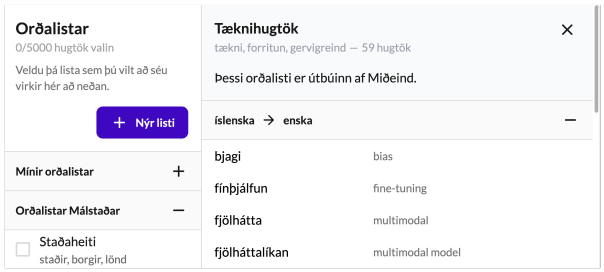

Figure 2: The user interface for the glossary integration in Erlendur (in Icelandic). The user can upload own glossaries or make use of glossaries shared by others in their Málstaður user group.

All the enhancement measures described in this section are intended to improve the translation capabilities beyond a baseline LLM translation. These measures are language-agnostic, unless otherwise stated. The modular nature of the system allows for most of these to be turned on and off. For the WMT25 submissions, all enhancements were used.

2.1 Model selection

While the general language quality of major languages such as English has seen diminishing returns in the latest generation of LLMs, the models still exhibit notable performance gaps when working with many lower-resource languages, and subtle or even significant differences can be noticed between different versions of the same model. We find, for example, that Claude 3.5 Sonnet (Anthropic, 2024) still surpasses other Claude model versions in Icelandic capabilities, even the more recent Claude 4 Sonnet/Opus (Anthropic, 2025). This is supported by Miðeind's Icelandic LLM leaderboard[4] (Miðeind, 2025) results, where Claude 3.5 Sonnet ranks highest of all models in the Claude model family in Icelandic language capabilities.

This motivated our selection of Claude 3.5 Sonnet as the underlying model in Erlendur at the time of system development, and this is the model used for all our WMT25 submissions. The system's API-based architecture makes it straightforward to substitute alternative models as Icelandic language capabilities advance.[5]

2.2 Preparatory analysis and lookup

For translation, the system only needs an input text and a target language; the source language is inferred if none is provided (optional parameters may be supplied). Once the input text is received, it is sent for analysis in a separate pass through the LLM. This is to gain a better understanding of the text, its style, subject domain, and general tone of voice. Key terms and named entities are extracted, and for English-to-Icelandic translations, a list of words is compiled for lookup in Ensk.is, an open-source dictionary[6] (Zoëga and Þórðarson, 2025). The LLM also identifies fixed expressions and idioms that may require special consideration in translation. The results are then prepared for incorporation into the main translation prompt along with the dictionary results provided through the dictionary API. The aim of this preparatory step is to provide the model with richer information on the type of text to translate, and to guide its focus on aspects of the text that require careful translation. This two-pass approach decouples analysis from generation, allowing for more explicit and targeted instructions in the final translation prompt. This step can be further enriched with more versatile tool use, such as dictionaries in other language pairs, or by providing explanations for complex concepts or idioms in the text.

2.3 Glossary integration

An important part of translations in a professional environment is consistent use of terminology. Busi-

[4] https://huggingface.co/spaces/mideind/icelandic-llm-leaderboard

[5] This architecture allowed us, after the WMT25 preliminary results were released (Kocmi et al., 2025b), to easily change the translation model to the high-scoring Gemini 2.5 Pro (https://deepmind.google/models/gemini/pro/), and now, at the time of publication, Gemini 2.5 Pro is the underlying model of Erlendur.

[6] https://ensk.is

nesses often compile glossaries of terms to ensure brand consistency, technical accuracy, and adherence to industry-specific language standards across all translated materials. Erlendur accepts custom glossaries (see Figure 2), where each term can be assigned a subject domain ("finance", "pharmaceutical", etc.) and even a special note on the usage of the term if needed. In the Erlendur editor, the glossary can be provided as a TSV file.

For longer glossaries, it is not feasible to inject them as a whole into the translation prompt, so we need to filter out the terms that appear in the input text. We also need to account for multi-word terms, and properly match those in the input text. Highly-inflected languages present unique challenges for term matching due to morphological variations. We developed a hybrid approach combining fuzzy matching and n-gram analysis, informed by the morphological characteristics of such languages.

The system generates n-grams of lengths 1 through `max_ngram_size` (determined by the longest glossary term) from the input text, then compares each n-gram against glossary terms of the same word count using a string-matching library (RapidFuzz[7]). The algorithm prioritizes longer n-grams first when selecting matches to favor multi-word term matches over shorter partial matches, and uses position tracking to avoid overlapping selections. Fuzzy matching is based on Levenshtein distance between text sequences, with a minimum similarity threshold of 0.75, established through iterative refinement during system development. The goal is to achieve high recall without generating an overly long list of candidate terms. We limit the list to a maximum of 50 terms per text chunk to translate, though this is adjustable, based on the chunk size (see Section 2.4). Once a list of term candidates has been compiled, the terms and their translations are incorporated into the translation prompt as an important resource for the LLM to follow when translating. This curated list of terms likely contains some false positives, but those can simply be ignored by the LLM. This methodology of fuzzy string matching of n-grams allows even inflected terms to be correctly matched, such as when the nominative term "sérstök áætlun" ("specific programme") matches the same term when it appears in the genitive case; "sérstakrar áætlunar". This ensures consistent term use throughout, even for longer documents.

[7] https://github.com/rapidfuzz/RapidFuzz

The glossary functionality is language-agnostic, and since the term matching module was developed to accommodate an inflectional language, it is flexible and should benefit many other morphologically rich languages. While in-house experiments indicate this, formal evaluations remain future work. Of note is that Erlendur placed first in the Terminology Translation Task, Track 2 (see Section 3.2), which tests terminology between Traditional Chinese and English, with Chinese being structurally and morphologically very different from Icelandic. In addition to providing terminological consistency, another benefit of injecting custom glossary terms when translating into lower-resource languages is that they can help fill the vocabulary gaps observed in LLMs for these languages.

2.4 Context-handling and translation

The translation prompt is an information-dense text with clear instructions on how to translate, along with the compiled analysis results, optional dictionary results and relevant glossary terms. An additional information string can be added, with special instructions or information about the text. For longer texts, whose translation might surpass the output token limit of the LLM in question, the system splits the text into fragments or chunks, and translates each chunk separately, while providing a snippet of the previous source text chunk as context to ensure text cohesion.

2.5 Post-processing

LLM-generated texts in Icelandic still contain ungrammatical sentences and made-up words. To remedy this, after translation, we run Icelandic target texts through our in-house grammatical error correction tool, Málfríður (Ingólfsdóttir et al., 2023). This helps catch ungrammatical sentences and correct them, mostly incorrect inflections or unconventional preposition use.

3 WMT submissions

We used Erlendur for both the general and the terminology shared tasks, with the same enhancements. The following sections describe the details of each submission.

3.1 General Machine Translation Task

In our submission, our aim was to use the features already present in Erlendur, without special handling for specific texts in the test set, to demonstrate

the robustness of the system and mimic realistic user behavior. The API offers the option of adding special instructions or information for the task at hand, as mentioned in Section 2.4. This option was used to relay some metadata from the test set, namely the domain, the `doc_id`, and a shortened version of the prompt string provided for each of the four focus domains (*news*, *speech*, *social*, *literary*).

One task-specific instruction was added: A considerable part of the test set data is in the first person (the *speech* domain in particular), and the speaker's gender is not always evident from the text. This calls for some decision-making when translating into Icelandic, to ensure gender agreement in the translation. Icelandic has inherent grammatical gender (masculine, feminine, or neuter), and adjectives change according to gender, so "I'm worried" translates into "Ég er áhyggjufull" (feminine) or "Ég er áhyggjufullur" (masculine), depending on the subject's gender. Instead of inferring the gender from the limited context of the source text, for the *speech* domain, we opted to ask the model to output the standard abbreviated gender notation, "Ég er áhyggjufull(ur)."

Another model-specific limitation is that Claude 3.5 Sonnet cannot produce Icelandic closing quotation marks (“); to remedy this we ask the model to instead output French quotation marks (guillemets, « ») and then we convert them to Icelandic ones („“) in post-processing.

For the translation, we used our standard glossary of place names and organizations in English, and their official Icelandic translations, compiled in-house. This glossary is available for use by Erlendur users.

3.2 Terminology Translation Task

The WMT25 Terminology Translation Task tests the inclusion of a given dictionary of terms when translating, to ensure correct and consistent terminology in specialized domains. This task has two tracks: Track 1 involves translating short text chunks from English into Russian/Spanish/German, and correctly incorporating term translations from a short list of terms that appear in the text. Track 2 better mimics real-life conditions, where the texts are corpus-level and the glossaries are considerably larger. The language pairs are English→Traditional Chinese and Traditional Chinese→English. In both tracks, there are three modes: no terminology, random terminology (the glossary terms are words randomly drawn from input texts) and proper terminology (domain-specific terminology).

In this task, we participated in both Track 1 and Track 2, making use of the native glossary functionality of Erlendur described in Section 2.3. The system's existing capabilities, particularly its efficient term matching and enforcement of terminological consistency, were sufficient to meet the task requirements without further modification. We, however, encountered an unexpected challenge in Track 2. The underlying model, Claude 3.5 Sonnet, refused to consistently generate Chinese translations of the test data, returning a "Content blocked" message. No prompt adjustments we tried could properly bypass this content filter, so we utilized our system's modular architecture to replace the underlying model with GPT 4.1 (OpenAI, 2025), which successfully processed the translations and has solid multilingual capabilities. As the source text itself was innocuous, we concluded that the filter was likely triggered by a policy related to the generation of the target language itself.

This is an example of unforeseen issues that may arise when using external, closed-source models over which the user has no control. It also underscores the value of a flexible system design that permits rapid adaptation, such as swapping the core LLM, to ensure robustness against external constraints.

3.3 WMT results

As briefly mentioned in the introduction, Erlendur ranked 3rd in the WMT25 General Machine Translation Task (Kocmi et al., 2025a) in English-Icelandic translations, with a human translator taking top place and Gemini 2.5 Pro taking second. GPT-4.1 was a close 4th, within the margin of statistical significance. Erlendur thus scored the highest out of all WMT25 participants for this language pair, even with a model that is relatively old and close to being deprecated (Claude Sonnet 3.5). Gemini 2.5 Pro, included for comparison, was the high-scoring model across most languages in the general MT task, also for Icelandic; this has been our cue to replace it as the underlying model in the current version of Erlendur. While Claude 3.5 Sonnet without enhancements was not evaluated in the task, Claude 4 ranked 8-10 for the language pair.

In the Terminology task, Erlendur was the only

system scoring in the top 5 in both Track 1 and Track 2, taking third place in Track 1 and first place in Track 2. This is a clear indicator that our terminology approach, developed with a focus on Icelandic and English, has a solid foundation that works for a range of languages, as different from Icelandic as Russian and Chinese. It also demonstrates its robustness with both longer input texts and its practical value in industry scenarios, where extensive glossaries are commonly used.

4 Conclusion

We have presented Erlendur, a multilingual LLM-based translation system that addresses the challenges of translating into lower-resource languages through a hybrid multi-stage approach. By combining preparatory text analysis, dictionary lookup, glossary integration, and grammatical error correction, the system demonstrates how targeted enhancements can significantly improve translation quality for languages like Icelandic without requiring model fine-tuning. The language-agnostic design of most system components makes this approach applicable to other lower-resource language pairs, while the modular architecture allows for flexible adaptation to different translation scenarios and future model improvements.

Limitations

The translation process of Erlendur is a hybrid pipeline that involves two passes through an LLM, some processing of terms, API lookup and a grammatical correction pass through a separate model, which means that the translation time is longer than in smaller models and simpler solutions. For a faster translation turnaround, each of the preprocessing steps can be included or skipped in the API. The bulk of the overall time, however, is spent on text generation by the model when producing the translation, while the preprocessing steps and dictionary lookups add minimal overhead. Concurrent handling makes time measurements of each step challenging, so while this information would be useful, it has not been reported in this work.

While we hypothesize that our approach, carefully developing a system to translate to and from a morphologically rich language, will benefit other languages of varying morphological complexity, this has not been confirmed with formal experiments.

References

Anthropic. 2024. Claude 3.5 Sonnet Model Card Addendum. https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf.

Anthropic. 2025. System Card: Claude Opus 4 & Claude Sonnet 4. https://www-cdn.anthropic.com/07b2a3f9902ee19fe39a36ca638e5ae987bc64dd.pdf.

Catherine Arnett and Benjamin Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6607–6623, Abu Dhabi, UAE. Association for Computational Linguistics.

Svanhvít Lilja Ingólfsdóttir, Pétur Ragnarsson, Haukur Jónsson, Haukur Símonarson, Vilhjálmur Þorsteinsson, and Vésteinn Snæbjarnarson. 2023. Byte-Level Grammatical Error Correction Using Synthetic and Curated Corpora. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7299–7316, Toronto, Canada. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica M. Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025a. Findings of the WMT25 General Machine Translation Shared Task: Time to Stop Evaluating on Easy Test Sets. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025b. Preliminary Ranking of WMT25 General Machine Translation Systems.

Miðeind. 2025. Icelandic LLM leaderboard. Accessed: 2025-7-1.

Miðeind. 2025. Málstaður.

OpenAI. 2025. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/. Accessed: 2025-08-11.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Kirill Semenov, Xu Huang, Vilém Zouhar, Nathaniel Berger, Dawei Zhu, Arturo Oncevay, and Pinzhen Chen. 2025. Findings of the WMT25 Terminology Translation Task: Terminology is Useful Especially for Good MTs. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Geir T. Zoëga and Sveinbjörn Þórðarson. 2025. Ensk.is.