

What a DRAG (It Is Being Small)

The AMI Submission to the WMT 2026 General Translation Shared Task

Steinþór Steingrímsson, Sveinbjörn Þórðarson, Jón Friðrik Daðason

The Árni Magnússon Institute for Icelandic Studies

Reykjavík, Iceland

steinthor.steingrimsson, sveinbjorn.thordarson,

jon.fridrik.dadason@arnastofnun.is

Abstract

We describe the AMI team’s submission to the WMT 2026 General Translation shared task, participating in the English→Icelandic translation direction. Building on our 2025 submission, which relied on a 3B-parameter model and substantial rule-based post-processing, this year we aim to reduce that dependency and rely more on the translation model itself. We continually pre-train a Qwen3-4B checkpoint on a mixture of Icelandic, English, code, math, and structured data, then fine-tune it with LoRA on a retrieval-augmented, domain-conditioned instruction-tuning dataset covering five domains. At both training and inference time, source sentences are enriched with dictionary entries and back-translated example translations similar to the sentence being translated, retrieved by a dedicated RAG server, and served through a self-healing inference client that retries and validates generations. We report COMET and chrF++ scores comparing this system against last year’s submission, a larger Llama-3.1-8B model adapted with the same recipe, and an instruction-tuned Qwen3-4B baseline, alongside ablations that isolate the contribution of each part of the retrieval-augmented prompt.

1 Introduction

Machine translation into Icelandic, a morphologically rich language with roughly 350,000 native speakers, remains a low-resource setting even as general-purpose large language models (LLMs) have become competitive translators for high-resource pairs. The AMI team’s 2025 WMT submission (Jasonarson and Steingrímsson, 2025) addressed this with a 3B-parameter fine-tuned model paired with substantial rule-based post-processing to correct systematic model errors after generation. For this year’s version (Kocmi et al., 2026), we continue experimenting with smaller models and investigate how far a smaller model can be pushed

by moving correction earlier in the pipeline: rather than post-editing a finished translation, we augment the model’s input with retrieved dictionary entries and similar example translations at both fine-tuning and inference time, and replace heavy post-hoc correction with a lightweight, on-the-fly validate-and-retry loop during decoding (Section 5). We refer to this combination—retrieving bilingual dictionary entries for terms appearing in the source and retrieving similar sentence pairs from a reference corpus, and injecting both into the translation prompt—as *dictionary retrieval-augmented generation* (DRAG), following the terminology of Zheng et al. (2024).

Concretely, we continually pre-train a Qwen3-4B (Qwen Team, 2025) checkpoint on a mixture of Icelandic, English, code, math, and structured-data text (Section 3.1), then fine-tune it with LoRA (Hu et al., 2022) on a retrieval-augmented, domain-conditioned instruction dataset spanning five domains (Section 3.2). We use LoRA rather than full fine-tuning as we are aiming for an approach that demands the least compute.

At both training and inference time, source sentences are enriched with entries from three bilingual dictionaries and with similar sentence pairs retrieved from a large back-translated reference corpus. The resulting adapter is served behind an OpenAI-compatible endpoint and queried through a client that validates each generation against a small set of rules and resamples on failure (Section 5).

We evaluate this system against our 2025 submission, an 8B-parameter Llama model adapted with the same recipe, and an instruction-tuned Qwen3-4B baseline that received no Icelandic-specific adaptation, alongside a set of ablations isolating the contribution of each component of the retrieval-augmented prompt (Section 6). We release the

fine-tuning dataset¹, retrieval pipeline², and model weights³ publicly.

2 Related Work

Adapting a multilingual base model to a target language via continued pre-training (CPT), rather than training from scratch, is now a common recipe for extending LLM coverage to lower-resource languages, see e.g. [Joshi et al. \(2025\)](#). When prompted with few-shot in-context examples, LLMs have been shown to be competitive translators ([Garcia et al., 2023](#)) without any translation-specific fine-tuning at all. Our system instead combines lightweight fine-tuning with retrieval-augmented prompting, aiming to get the benefit of both.

Low- and medium-resource languages often lack sufficient data to pre-train broadly capable models from scratch and are frequently underrepresented in multilingual pre-training corpora, resulting in weaker performance than for high-resource languages. A promising alternative is to adapt an existing model to the target language through CPT, which has yielded substantial improvements for languages such as Basque ([Corral et al., 2025](#)) and Estonian ([Dorkin et al., 2026](#)). However, training predominantly on target-language text risks catastrophic forgetting of capabilities inherited from the base model, degrading performance in other languages and domains ([Zhao et al., 2024](#); [Elhady et al., 2025](#)). Replaying English data during CPT mitigates this degradation ([Elhady et al., 2025](#)), and recent language-adaptation efforts accordingly mix target-language text with English ([Fujii et al., 2024](#); [Joshi et al., 2025](#)) and, sometimes, with data such as code and mathematics ([Dorkin et al., 2026](#)).

Selecting good in-context examples for few-shot prompting has been shown to materially affect LLM translation quality ([Agrawal et al., 2023](#)), and retrieval over a reference corpus of similar translations is one way to construct such examples automatically ([Lee et al., 2025](#)). We adopt this idea for both fine-tuning and inference (see Section 3.2.3), retrieving over a corpus built with back-translation.

Dictionary-augmented prompting has already shown some promise when it comes to English–

Icelandic translation. [Ingólfssdóttir et al. \(2025\)](#) describe Erlendur, a multi-stage pipeline that uses a preparatory LLM pass to extract key terms from the source text, looks these up in the Ensk.is dictionary (see Section 3.2.2) and in user-supplied glossaries using inflection-aware fuzzy n-gram matching, and injects the results into an LLM translation prompt.

An alternative to injecting retrieved context directly into the prompt is to expose the dictionary and example retrieval as tools the model can call at its own discretion, e.g. via Model Context Protocol (MCP).⁴ We considered this approach but decided on using prompt injection for our submitted system: it guarantees that relevant context is present in every prompt regardless of whether the model would have initiated a lookup, at the cost of always paying the retrieval latency and prompt-length overhead even when a given sentence needed no help.

Icelandic has been a target language in the WMT General Translation shared task in recent years ([Kocmi et al., 2024, 2025](#)). We compare directly against our own 2025 submission ([Jasonarson and Steingrímsson, 2025](#)) in Section 6.

3 Data and Augmentation Data

For pre-training and fine-tuning we use a wide array of available datasets. In Sections 3.1 and 3.2 we give an overview of these datasets and how they were used.

3.1 Pre-training Data

Our CPT corpus contained approximately 7.39B tokens, comprising Icelandic (63.0%), English text (27.0%), Python code (4.0%), math (3.5%), and structured data (2.5%). The Icelandic subset consisted of 4.65B tokens from a filtered and deduplicated version ([Daðason and Steingrímsson, 2026](#)) of the Icelandic Gigaword Corpus (IGC) ([Steingrímsson et al., 2018](#); [Barkarson et al., 2022](#)). For English-language text, we sampled 1.79B tokens from a deduplicated version of FineWeb-Edu ([Penedo et al., 2024](#)) and 199M tokens from Cosmopedia v2 ([Ben Allal et al., 2024](#)). We sampled 295M tokens from the Python subset of the StarCoder dataset ([Li et al., 2023](#)), and 259M tokens of mathematical content from FineMath-4plus ([Allal et al., 2025](#)). Finally, we extracted 185M tokens from English-language articles in Structured Wikipedia ([Wikimedia Enterprise and Wikimedia Foundation, 2026](#)), with articles in plain text with

¹https://github.com/stofnun-arna-magnussonar/WMT2026_finetuning_dataset

²<https://github.com/stofnun-arna-magnussonar/WMT2026>

³<https://huggingface.co/arnastofnun/Qwen3-4B-wmt26-AMI-en-is> and <https://huggingface.co/arnastofnun/Llama-3.1-8B-wmt26-AMI-en-is>

⁴<https://modelcontextprotocol.io/specification>

JSON-formatted tables. All token counts were measured using the Qwen3 tokenizer.

3.2 Fine-tuning Data

To adapt our continually pre-trained model for the translation task and teach it to effectively utilize injected context, we construct a multifaceted fine-tuning dataset. Because our system uses retrieval-augmented generation (RAG) at inference time, the fine-tuning data must teach the model how to adhere to domain-specific instructions, apply retrieved dictionary constraints, and mimic structural few-shot examples. To achieve this, our dataset relies on three distinct components: a core set of domain-annotated parallel sentences, a comprehensive collection of bilingual dictionary entries, and a large-scale back-translated reference corpus. We detail each of these resources below.

3.2.1 Training examples

Training examples for the *news* domain are drawn from the official WMT 2024 English–Icelandic test set, split into paragraph-level segments; *general*, *software*, and *speech* domains were synthetically generated with Gemini 3.1 Pro from seed prompts (Appendix B); *social* examples are Gemini-generated from seed prompts drawn from WMT 2025 general translation task test data (Kocmi et al., 2025). In total we used 4,996 examples across the five domains (1,000 general, 996 news, 1,000 social, 1,000 software, 1,000 speech), but although the number of examples is almost the same for all domains, note that the length of the examples differs somewhat, as seen in Table 1. The fine-tuning dataset is available on GitHub.⁵

Domain	# Ex.	Source	Avg. Words
General	1,000	Gemini 3.1 Pro	20.9
News	996	WMT24 test set	32.5
Social	1,000	Gemini 3.1 Pro	23.8
Software	1,000	Gemini 3.1 Pro	49.9
Speech	1,000	Gemini 3.1 Pro	32.5
Total	4,996		–

Table 1: Fine-tuning data by domain.

3.2.2 Dictionary lookups

We assembled three English–Icelandic dictionary resources: *Ensk.is*, an open-source English–Icelandic dictionary (40,547 headword entries)

⁵https://github.com/stofnun-arna-magnussonar/WMT26_AMI

(Zoëga and Þórðarson, 2026); a general-purpose glossary generated semi-automatically and then manually error-checked (94,061 entries) (Steingrímsson et al., 2022; Steingrímsson et al., 2021), and *Tölvuorðasafnið*, a computing terminology lexicon containing 7,171 entries (Guðmundsdóttir et al., 2026). Together, these give 141,779 dictionary entries available for lookup at both fine-tuning and inference time. The retrieval and scoring mechanism that selects which entries are injected into a given prompt is described in Section 4.2.2.

3.2.3 Back-translated few-shot examples

To provide naturalistic few-shot translation examples beyond what our fine-tuning data covers, we used a 6-bit quantized version of TranslateGemma 27B⁶ (Finkelstein et al., 2026) to build a large reference corpus by back-translating (Sennrich et al., 2016) monolingual Icelandic text from the Icelandic Gigaword Corpus (IGC). We generated an English translation for a large sample of sentences, yielding synthetic English–Icelandic pairs in which the Icelandic side is authentic, naturally occurring text and the English side is machine-generated. We filtered candidate pairs on perplexity calculated by a KenLM (Heafield, 2011) model trained on the 2024 version of the IGC (Steingrímsson, 2026) and automatic language identification using a custom langdetect⁷ profile for Icelandic⁸ to remove poor translations and off-language content, retaining approximately 106.5 million sentence pairs (Steingrímsson and Hafsteinsson, 2026a,b). This corpus is queried at both fine-tuning and inference time to retrieve similar example translations for a given source sentence, via the FTS5 and MinHash/LSH indices described in Section 4.2.1.

4 Building the System

Our submitted translation system is a lightweight, retrieval-augmented pipeline centered around a 4B-parameter language model. Building the system involved a two-stage training process coupled with the development of a dedicated context-retrieval engine. First, we adapted the base model to the Icelandic language via efficient, small-batch continued pre-training. Second, we fine-tuned the model using Low-Rank Adaptation (LoRA) to teach it

⁶https://huggingface.co/mradermacher/translategemma-27b-it-i1-GGUF:Q6_K

⁷<https://pypi.org/project/langdetect/>

⁸https://github.com/steinst/langdetect_profiles

how to follow domain-specific instructions and leverage dynamically injected context. To supply this context, we built a standalone Retrieval-Augmented Generation (RAG) server that scores and retrieves relevant bilingual dictionary entries and back-translated few-shot examples on the fly. The following subsections detail the pre-training and fine-tuning procedures, the architecture of the RAG server, and the mechanisms used to filter, score, and assemble the final context-enriched prompts.

4.1 Pre-training

We adapted the Qwen3-4B base model⁹ (Qwen Team, 2025) to Icelandic through continued pre-training. We followed the small-batch training approach of Marek et al. (2025), who found that, when Adam’s second-moment decay is scaled to maintain a fixed token half-life, small-batch training is stable, more robust to hyperparameter choices, and achieves equal or better per-FLOP performance than larger-batch training. We used a sequence length of 8,192 and a batch size of 1, and set AdamW’s second-moment decay rate to $\beta_2 = 0.9994323$, corresponding to a token half-life of 10M tokens. We used a WSD learning rate schedule, with a linear warmup over the first 2,000 steps to a peak learning rate of $3e-5$, and a linear decay over the final 10% of steps. We trained on a single NVIDIA RTX PRO 6000 Blackwell GPU for 24 hours and for the downstream experiments, we used the checkpoint saved after 75,000 steps, corresponding to 614M training tokens. Training up to this checkpoint took approximately 23 hours.

4.2 Fine-tuning

We fine-tuned the continually pre-trained checkpoint described above using low-rank adaptation (LoRA) (Hu et al., 2022) as implemented in Unsloth¹⁰ (Daniel Han and team, 2023). We used LoRA rank $r = 16$, $\alpha = 16$, dropout 0, no bias term, applied to all attention and MLP projection matrices (q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj). We trained for 3 epochs with a per-device batch size of 2 and 4 gradient accumulation steps (effective batch size 8), a peak learning rate of $2e-4$ with a linear schedule and 10 warmup steps, the 8-bit AdamW optimizer, weight decay 0.01, and bf16 mixed precision.

⁹<https://hf.co/Qwen/Qwen3-4B-Base>

¹⁰<https://github.com/unslothai/unsloth>

Each training example pairs a domain-specific system prompt (*social*, *speech*, *news*, *software*, or *general*) with a user turn containing retrieved dictionary entries and example translations for the source sentence (retrieval and prompt construction detailed in Section 4.2.3), and an assistant turn containing the reference Icelandic translation. Examples were rendered through Qwen3’s native ChatML-style chat template, with loss computed over the assistant turn only (see Appendix A). We trained on the 4,996 examples described in Section 3.2, with dictionary lookups added to the prompt as well as translations of similar sentences.

4.2.1 Back-translation RAG server

Prompt construction is handled by a single persistent FastAPI service (Ramírez, 2026) that keeps three bilingual dictionaries, a sentence-retrieval index, and a cross-encoder reranker resident in memory, run as a standalone Uvicorn (Trylesinski and Christie, 2026) process and queried over HTTP by both the dataset-building and inference clients. It performs one warm-up retrieval query on startup, then exposes a context endpoint returning a ranked, capped dictionary block and example block as separate fields (the caller assembles the final prompt, Section 4.2.3); an alternative single-call endpoint that returns one pre-assembled prompt exists but is unused by either pipeline.

The example block draws on the large back-translated Icelandic reference corpus (see Section 3.2.3) via two indices: an SQLite FTS5 full-text index over (source, target) pairs, queried by n-gram phrase matching; and a MinHash locality-sensitive-hashing (LSH) index (Broder, 1997) for approximate nearest-neighbor retrieval. Each record’s English side is shingled into bigrams, hashed into a 64-permutation MinHash signature, split into 16 bands of 4 rows for bucketing, and candidates are rescored by exact Jaccard similarity; matched records are fetched by direct byte-offset seek via a precomputed offset map.

4.2.2 Dictionary lookup mechanism

The three static English–Icelandic resources described in Section 3.2.2 are loaded at startup: a general English–Icelandic dictionary, a semi-automatically generated, manually-checked glossary, and an EN-IS computing terminology lexicon, each a flat mapping from English word/phrase to Icelandic candidate(s). Multi-word entries are indexed by token sequence so the longest matching

phrase in a source sentence is preferred over its component words. Single tokens are looked up directly, on failure, under WordNet-derived lemma candidates (noun/verb/adjective/adverb) (Miller, 1995) via NLTK (Bird and Loper, 2004), so inflected forms resolve to their dictionary headword; English stopwords are skipped.

4.2.3 Prompt Construction

The source text is first segmented into sentences (HTML markup and JSON-array-of-strings inputs are normalized into plain segments), which drives both dictionary lookup and per-sentence example retrieval.

Dictionary block. For long sentences a very long list of dictionary entries can be found, some of them for common words that the LLM should translate correctly without the help of lookups. In order to reduce long lists of entries to something more manageable, we built an algorithm to find words that are more likely to be rare. To do this, hits are scored. Token rarity weight is $\max(0.1, 7.0 - \text{zipf}(\text{token}))$ (Speer, 2022). General-dictionary single-token hits are only considered if the token’s Zipf frequency is ≤ 4.5 (the computing lexicon is exempt from this gate). Score is (in-sentence token frequency) $\times$ rarity for single tokens, or (rarest component word’s rarity) $+ 1.5$ for phrase matches; computing-lexicon hits add a further $+2.0$. The top 15 scoring entries survive (words and phrases share this cap); the block is omitted if none are found.

Example block. Per sentence, up to 15 LSH neighbors are retrieved and filtered to pairs with ≥ 5 words on both sides, keeping the top 5; up to 5 FTS5 matches are found by querying n-grams from 7-grams down to 4-grams (falling back to single non-stopword tokens), pooled to 50 candidates before random subsampling to 5. All candidates are scored against their sentence with a cross-encoder¹¹ (Reimers and Gurevych, 2019) and merged into one pool across the whole input, keeping the best score per unique pair. The pool is filtered to score ≥ -11.0 ; if fewer than 2 candidates survive, no examples are emitted. Otherwise a top band of 6–24 candidates (size drawn uniformly at random) is formed, and 2–8 examples (likewise randomly drawn, capped by band size) are sampled without replacement and shown best-score-first, so the injected example count varies per prompt. This was

done so that when fine-tuning, the model learns to translate regardless of the number of few-shot examples, as in some cases they may be none, while in others very many sentence pairs can be found.

Assembly. The dictionary block (if any) and example block (if any) are each wrapped in a fixed instruction and concatenated with a final translation instruction containing the raw source sentence; a missing block is omitted, not left empty. The fine-tuning and translation prompt and chat template are given in Appendix A. The fine-tuning dataset is available in our GitHub repo.

5 Inference

At inference time, source sentences are passed through the identical retrieval and prompt-construction pipeline used to build the fine-tuning data, and the fine-tuned adapter is served behind an OpenAI-compatible endpoint.

5.1 Rule-based re-translation

Generations are checked against a fixed set of validation rules before being accepted. A translation is rejected if it: leaks the dictionary- or example-block instruction text verbatim; contains an English-prompt hallucination (e.g. an echoed “English:” label); is empty; repeats the same 5-word sequence five or more times; introduces more paragraph breaks than the source; or, on segments of at least 15 words, has a translation-to-source word-count ratio below 0.70 or above 1.50, or copies more than 50% of its content words directly from the source. Length-ratio and source-copy checks are skipped on shorter segments, where they are not statistically meaningful.

A rejected or truncated (max-token-limited) generation is resampled: the first 11 retries linearly increase temperature from 0.25 to 0.75; if that fails to produce a passing output, five further retries hold temperature at 0.27 and instead sweep the sampling parameters (top_p , top_k , min_p) from a tight to a wide cutoff, on the assumption that a low-temperature failure is a distribution-shape problem rather than one that needs more randomness. The 16th and final attempt is accepted unconditionally regardless of validation outcome, so every input is guaranteed an output.

6 Results

In order to evaluate our system, we split the test set from the WMT 2025 general translation task

¹¹<https://hf.co/cross-encoder/ms-marco-MiniLM-L6-v2>

Dictionaries	Few-shot	n	COMET	chrF++
✓	✓	332	0.7792 ^a	45.88 ^a
✓	×	332	0.7772 ^a	45.90 ^a
×	✓	332	0.7695 ^b	45.19 ^b
×	×	332	0.7697 ^b	45.19 ^b

Table 2: COMET and chrF++ scores by ablation (dictionary/examples present or not; all domains pooled, $n = 332$). Scores sharing a superscript letter for a given metric are not significantly different from each other on that metric (paired bootstrap resampling, $n_{\text{samples}} = 1000$, seed=12345); different letters indicate $p < 0.05$.

(Kocmi et al., 2025) into paragraphs, translated these with our system using different prompting approaches and calculated COMET¹² (Rei et al., 2022) and chrF++¹³ scores (Popović, 2017).

We started by doing an ablation study, where we prepared the prompt in four different ways: including the dictionary lookups, including few-shot examples, including both or neither. As shown in Table 2 we get the best results if we include the dictionary lookups, but there is not a statistically significant difference if we add the few-shot examples.

We decide to stick with using both few-shot examples and the dictionaries for translating our WMT 2026 submission. While the automatic scoring shows a clear positive advantage in using the dictionaries, it doesn’t indicate a statistically significant difference when using the few-shot examples. We decide to use the few-shots examples anyway, as such examples have been shown to increase LLM translation quality, see e.g. Agrawal et al. (2023), and the manual evaluation might be more sensitive to finer nuances. On the other hand, if a large back-translated corpus does not exist, our results would not justify one being built for this purpose only.

While our submitted model this year is of similar size to the one submitted last year, the translation pipeline is somewhat easier to run. The post-processing part was quite cumbersome last year, making the system complicated to set up and run in practice. But do we lose quality by skipping the post-processing? In Table 3 we compare our system to our previous submission, as well as a few other available models, and we do indeed find that the system submitted by the AMI team this year is

¹²COMET signature: COMETmodel:Unbabel/wmt22-comet-dalcomet:2.2.7

¹³chrF++ signature: nrefs:1 | bs:1000 | seed:12345 | case:mixed | eff:yes | nc:6 | nw:2 | space:no | version:2.6.0

System/model	COMET	chrF++
AMI 2025	0.7950	48.70
Llama8B+CPT+LoRA	0.7895	46.98
AMI 2026 (4B)	0.7792	45.88
Qwen3-4B-Instruct-2507	0.4569	30.87

Table 3: Comparison of COMET and chrF++ scores for our system and others.

inferior to the submission from last year. By cutting out the post-processing we do seem to lose some quality, although other things may affect the results too, such as the short CPT-process for this model, or the fine-tuning process which was much simpler last year when we fine-tuned it simply by showing it very many examples of parallel sentences, and always used the same short prompt. What is gained on the other hand is a greater overall simplicity of the system.

We also compared our model to another model we did CPT on, an 8B-parameter Llama 3 model (Llama Team, 2024), and then trained a LoRA adapter in exactly the same way as we did with our submitted model. The 8B-parameter Llama 3 was trained on Icelandic, English and Polish, 25 billion tokens in total, 40% Icelandic, 40% Polish and 20% English. While the COMET and chrF++ scores are higher for this model than the submitted model, upon manual inspection of the outputs, we found that this model was much more prone to hallucinations and going off track, although the fluency of the Icelandic texts looked better than in our submitted model. As the submitted model seemed to be more likely to follow instructions correctly we decided to go with that.

Finally, we compared our model to Qwen3-4B-Instruct-2507, an instruction-tuned variant of the base model we continued pre-training on. While this comparison does not isolate the impact of continued pre-training and fine-tuning, it represents a realistic alternative when no language-specific training is performed. As expected, our adapted model substantially outperforms this baseline, which produces low-quality Icelandic output and often fails to follow the domain prompts.

7 Conclusions and Future Work

We have evaluated DRAG only for a small language model. Whether dictionary retrieval also improves English–Icelandic translation for state-of-the-art LLMs, which already handle the language pair reasonably well, remains to be measured.

Our dictionary retrieval draws solely on English-keyed resources. The DRAG pipeline could potentially be complemented by additional data, such as *Íslensk-ensk orðabók*,¹⁴ an Icelandic–English dictionary that could be queried in reverse. Matching for English source words in IS-EN definitions might bring to the surface Icelandic vocabulary and translations that our English-keyed resources do not contain.

The Qwen3 tokenizer is inefficient for Icelandic, with each token representing only 2.22 characters on average in the Icelandic subset of our CPT corpus. This reduces the effective context length in terms of Icelandic text and increases the number of tokens, and consequently the computation, required to process a given amount of text. In future work, we intend to investigate longer CPT runs together with vocabulary extension to improve the model’s tokenization efficiency for Icelandic.

And finally, it may very well be so that when fine-tuning an LLM to be effective at machine translation for a low-resource language, the small-sized model we are working with is not big enough to produce reasonable quality, even after CPT, especially when trying to make it follow long and complicated prompts containing a lot of information. We leave these questions for further work.

References

- Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. *In-context Examples Selection for Machine Translation*. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.
- Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, and 3 others. 2025. *SmolLM2: When Smol Goes Big – Data-Centric Training of a Small Language Model*. *Preprint*, arXiv:2502.02737.
- Starkaður Barkarson, Steinþór Steingrímsson, and Hildur Hafsteinsdóttir. 2022. *Evolving Large Text Corpora: Four Versions of the Icelandic Gigaword Corpus*. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2371–2381, Marseille, France. European Language Resources Association.
- Loubna Ben Allal, Anton Lozhkov, Guilherme Penedo, Thomas Wolf, and Leandro von Werra. 2024. *Cosmopedia*.
- Steven Bird and Edward Loper. 2004. *NLTK: The Natural Language Toolkit*. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.
- A.Z. Broder. 1997. *On the resemblance and containment of documents*. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, pages 21–29.
- Ander Corral, Ixak Sarasua Antero, and Xabier Saralegi. 2025. *Pipeline Analysis for Developing Instruct LLMs in Low-Resource Languages: A Case Study on Basque*. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12636–12655, Albuquerque, New Mexico. Association for Computational Linguistics.
- Michael Han Daniel Han and Unsloth team. 2023. *Unsloth*.
- Jón Friðrik Daðason and Steinþór Steingrímsson. 2026. *Optimized for AI: Curating the Icelandic Gigaword Corpus for Stable LLM Training*. In *Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora*, pages 49–56, Palma, Mallorca, Spain. European Language Resources Association (ELRA).
- Aleksei Dorkin, Taïdo Purason, Emil Kalbaliyev, Hele-Andra Kuulmets, Marii Ojastu, Mark Fišel, Tanel Alumäe, Eleri Aedmaa, Krister Kruusmaa, and Kairit Sirts. 2026. *EstLLM: Enhancing Estonian Capabilities in Multilingual LLMs via Continued Pretraining and Post-Training*. *Preprint*, arXiv:2603.02041.
- Ahmed Elhady, Eneko Agirre, and Mikel Artetxe. 2025. *Emergent Abilities of Large Language Models under Continued Pre-training for Language Adaptation*. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32174–32186, Vienna, Austria. Association for Computational Linguistics.
- Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, and 2 others. 2026. *TranslateGemma Technical Report*. *Preprint*, arXiv:2601.09012.
- Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. *Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities*. In *First Conference on Language Modeling*.

¹⁴<https://enska.arnastofnun.is/>

- Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Melvin Johnson, and Orhan Firat. 2023. [The Unreasonable Effectiveness of Few-shot Learning for Machine Translation](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10867–10878. PMLR.
- Arnheiður Guðmundsdóttir, Ásta Kristín Hauksdóttir Wíium, Guðmundur Friðrik Magnússon, Guðmundur Hafsteinn Viðarsson, Hafsteinn Einarsson, Linda Björk Bergsveinsdóttir, Ólafur Andri Ragnarsson, Sigurður Emil Pálsson, Stefán Ólafsson, Torfi Þórhallsson, Þorkell Guðmundsson, Baldur Jónsson, Sigrún Helgadóttir, Þorsteinn Sæmundsson, Örn S. Kaldalóns, and Steinþór Steingrímsson. 2026. [Tölvuorðasafn 2026.09](#). CLARIN-IS.
- Kenneth Heafield. 2011. [KenLM: Faster and smaller language model queries](#). In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-Rank Adaptation of Large Language Models](#). In *International Conference on Learning Representations*.
- Svanhvít Lilja Ingólfssdóttir, Haukur Páll Jónsson, Kári Steinn Aðalsteinsson, Róbert Fjölfnir Birkisson, Sveinbjörn Þórðarson, and Þorvaldur Páll Helgason. 2025. [Miðeind at WMT25 General Machine Translation Task and Terminology Translation Task](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 577–582, Suzhou, China. Association for Computational Linguistics.
- Atli Jasonarson and Steinthor Steingrímsson. 2025. [AMI at WMT25 General Translation Task: How Low Can We Go? Finetuning Lightweight Llama models for Low Resource Machine Translation](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 695–704, Suzhou, China. Association for Computational Linguistics.
- Raviraj Joshi, Kanishk Singla, Anusha Kamath, Raunak Kalani, Rakesh Paul, Utkarsh Vaidya, Sanjay Singh Chauhan, Niranjana Wartikar, and Eileen Long. 2025. [Adapting Multilingual LLMs to Low-Resource Languages using Continued Pre-training and Synthetic Corpus: A Case Study for Hindi LLMs](#). In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 50–57, Abu Dhabi. 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 Lakouagna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. [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*, pages 355–413, Suzhou, China. Association for Computational Linguistics.
- Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Sergey Dukanov, Anton Dvorkovich, Mark Fishel, Markus Freitag, Samuel Frontull, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Stan Kharevich, Philipp Koehn, Zheng Li, Jean Maillard, Christof Monz, Alexander Murauski, and 12 others. 2026. [Findings of the WMT26 general machine translation shared task: Contrastive dynamic human evaluation at scale](#). In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. [Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.
- Dohyun Lee, Seungil Chad Lee, Chanwoo Yang, Yujin Baek, and Jaegul Choo. 2025. [Exploring In-context Example Generation for Machine Translation](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26554–26568, Vienna, Austria. Association for Computational Linguistics.
- Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Joel Lamy-Poirier, Joao Monteiro, Nicolas Gontier, Ming-Ho Yee, and 39 others. 2023. [StarCoder: may the source be with you!](#) *Transactions on Machine Learning Research*.
- Llama Team. 2024. [The Llama 3 Herd of Models](#). Preprint, arXiv:2407.21783.
- Martin Marek, Sanae Lotfi, Aditya Somasundaram, Andrew Wilson, and Micah Goldblum. 2025. [Small Batch Size Training for Language Models: When Vanilla SGD Works, and Why Gradient Accumulation is Wasteful](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 148837–148862. Curran Associates, Inc.
- George A. Miller. 1995. [WordNet: A Lexical Database for English](#). *Commun. ACM*, 38(11):39–41.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024.

- The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. In *Advances in Neural Information Processing Systems*, volume 37, pages 30811–30849. Curran Associates, Inc.
- Maja Popović. 2017. *chrF++: words helping character n-grams*. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- Qwen Team. 2025. *Qwen3 Technical Report*. Preprint, arXiv:2505.09388.
- Sebastián Ramírez. 2026. *FastAPI*.
- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. *COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task*. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. *Improving Neural Machine Translation Models with Monolingual Data*. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.
- Robyn Speer. 2022. *rspeer/wordfreq: v3.0*.
- Steinþór Steingrímsson and Hinrik Hafsteinsson. 2026a. *Igc-2026-en1 – english translations 1*. CLARIN-IS.
- Steinþór Steingrímsson and Hinrik Hafsteinsson. 2026b. *Igc-2026-en2 – english translations 2*. CLARIN-IS.
- Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. *Risamálheild: A Very Large Icelandic Text Corpus*. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018*, pages 4361–4366, Miyazaki, Japan.
- Steinþór Steingrímsson, Luke O’Brien, Finnur Ingimundarson, Hrafn Loftsson, and Andy Way. 2022. *Compiling a Highly Accurate Bilingual Lexicon by Combining Different Approaches*. In *Proceedings of Globalex Workshop on Linked Lexicography within the 13th Language Resources and Evaluation Conference*, pages 32–41, Marseille, France. European Language Resources Association.
- Steinþór Steingrímsson, Luke James Obrien, Finnur Ágúst Ingimundarson, Árni Davíð Magnússon, Þórdís Dröfn Andréðóttir, and Inga Guðrún Eiríksdóttir. 2021. *English-Icelandic/Icelandic-English glossary 21.09*. CLARIN-IS.
- Steinþór Steingrímsson. 2026. *IGC-2024 5-gram: Icelandic KenLM language model trained on the Icelandic gigaword corpus (2024)*. CLARIN-IS.
- Marcelo Trylesinski and Tom Christie. 2026. *Uvicorn*.
- Wikimedia Enterprise and Wikimedia Foundation. 2026. *Structured Contents Wikipedia*.
- Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. *LLaMA Beyond English: An Empirical Study on Language Capability Transfer*. Preprint, arXiv:2401.01055.
- Jiawei Zheng, Hanghai Hong, Feiyan Liu, Xiaoli Wang, Jingsong Su, Yonggui Liang, and Shikai Wu. 2024. *Fine-tuning Large Language Models for Domain-specific Machine Translation*. Preprint, arXiv:2402.15061. Version 2, titled “DragFT: Adapting Large Language Models with Dictionary and Retrieval Augmented Fine-tuning for Domain-specific Machine Translation”.
- Geir T. Zoëga and Sveinbjörn Þórðarson. 2026. Ensk.is. <https://ensk.is>. Accessed: 2026-08-03. Processed dataset used in this work: https://huggingface.co/datasets/arnastofnun/ensk_is_parsed_wmt2026.

A Translation Prompt and Chat Template

Each prompt is assembled from up to three parts, each emitted only if the corresponding retrieval step returned results:

The following are reference translations retrieved for words and phrases that appear in the text below. Use them to guide your word and term choices, selecting the option that best fits the context and inflecting it (case, number, gender, tense) as Icelandic grammar requires. They are aids, not constraints: ignore any that do not fit, and translate words not listed here as usual.

{dictionary_block}

The following are example sentence translations from a corpus that contains similar phrasing or contextual themes. Use them as structural few-shot references:

{examples_block}

Translate the following English text into Icelandic. Only return the best possible Icelandic translation, without any explanations or notes.

{source_text}

The assembled prompt is rendered into the model’s chat format as:

```
<|im_start|>system
{domain system prompt}<|im_end|>
<|im_start|>user
{dictionary block (optional)}

{examples block (optional)}
```

Translate the following English text into Icelandic. Only return the best possible Icelandic translation, without any explanations or notes.

```
{source_text}<|im_end|>
<|im_start|>assistant
{translation}<|im_end|>
```

B Synthetic Data Generation Prompts

Training examples for the *general*, *social*, *software*, and *speech* domains were synthetically generated with Gemini 3.1 Pro (*gemini-3.1-pro-preview*), using structured output constrained to a fixed JSON schema (source text, target text, document/paragraph IDs, and domain metadata) and generated in batches of 5–10 examples per API call. The *news* domain was not Gemini-generated; it is drawn from the official WMT24 English–Icelandic test set (Section 3.2). The base prompt for each domain is reproduced below; batch-specific instructions (starting paragraph ID, exact batch size) were appended automatically at generation time.

B.1 General domain (temperature 0.9)

Generate a batch of highly diverse, general-purpose text in English and their natural translations into Icelandic. The topics should be completely arbitrary—ranging from scientific facts and philosophical musings to mundane tasks, historical trivia, or creative storytelling. CRITICAL CONSTRAINT: Each source text must be between 7 and 40 words long. Ensure the Icelandic translations are natural, fluent, and stylistically appropriate for the content.

B.2 Social media domain (temperature 1.0)

Generate a batch of highly diverse social media posts written in English and their natural translations into Icelandic. The posts should cover a wide variety of topics: hobbies, tech, travel frustrations, gym milestones, pet antics, cooking failures, dating, work, or daily observations. Use realistic

social media formatting, including emojis, casual phrasing, slang, internet acronyms, and appropriate cultural adaptations in Icelandic (e.g., translating memes/slang accurately rather than literally).

B.3 Software domain (temperature 0.4)

Unlike the other three domains, the software prompt is constructed per-batch: for each example in the batch, 1–6 English computing terms are sampled at random from the computing-terminology lexicon (Section 3.2.2), and the model is required to use both the English terms in the source text and their dictionary-specified Icelandic translations in the target text, enforcing terminological consistency with the dictionary the same system also uses at inference time. The shared instruction, with one example’s term constraint shown:

You are an expert bilingual technical writer and translator (English to Icelandic). Generate a batch of highly diverse texts focused on software and technology. For each example, you MUST naturally incorporate the specific English terms provided into the source text, and you MUST use the exact mapped Icelandic terms in the translated text.

Example 1: - Context: Write 1–3 short paragraphs (separated by \n or \n\n). It can be a software description, tech news, or a general webpage mentioning tech.
- Required English terms to include in src_text: [...] - Required Icelandic terms to include in tgt_text: [...] - Ensure the Icelandic translation strictly uses these exact translated terms.

B.4 Speech domain (temperature 1.0)

Generate a batch of transcribed speech examples in English and their natural translations into Icelandic. The English text should read exactly like a raw transcript from a YouTube video, podcast, or interview. It MUST include characteristics of unscripted spoken language: filler words (um, uh, you know), run-on sentences, abrupt topic changes, self-corrections, and conversational phrasing. Topics should range from makeup tutorials, gaming let’s plays, academic lectures, random vlogs, to fitness coaching. The Icelandic translation must accurately convey this messy, spoken tone rather than reading like formal text.

Domain	System prompt
General	You are a professional Icelandic translator, tasked with providing translations suitable for use in Icelandic (isl_Latn). Your goal is to accurately convey the meaning and nuances of the original text while adhering to Icelandic grammar and vocabulary and ensuring that the translation is natural. Produce only the Icelandic translation, without any additional explanations or commentary. Maintain the formatting of the original source text. Please translate the following text into Icelandic (isl_Latn):
News	... The original text is a news article. Ensure the translation is formal and consistent with journalistic standards. Produce only the Icelandic translation, without any additional explanations or commentary. Maintain the HTML formatting of the original source text. Please translate the following text into Icelandic (isl_Latn):
Social	... The original text is user-generated content from a social media platform and is accompanied by anonymized screenshot images that provide visual context. Use the screenshot context to resolve ambiguity, but translate only the text provided for translation. Do not reproduce spelling mistakes. Reproduce marks of expressiveness that communicate meaningful intent (e.g. enthusiasm through capitalisation or elongation) in a way that is natural in Icelandic. Copy URLs and user handles directly rather than translating them. However, translate hashtags as appropriate for the translation to be natural for social media text. Follow the punctuation of the source text as best as possible. Additional punctuation should be added only if not doing so would seriously alter the comprehension of the text. Translate text in an informal style, like close friends talking, even if it changes the original tone. Produce only the Icelandic translation, without any additional explanations or commentary. Maintain the HTML formatting of the original source text. Please translate the following text into Icelandic (isl_Latn):
Speech	... The original text is automatically transcribed from spoken language and can contain errors. Maintain the flow and colloquial style of the speaker in the translation. Do not include non-linguistic sounds (e.g. laughter, groans, hesitation sounds, etc.), but do include interjections. If a word is interrupted, either guess the full word if possible or otherwise omit it. Keep foreign words as they are when translating. Produce only the Icelandic translation, without any additional explanations or commentary. Output the text such that each sentence is on a separate line. Please translate the following text into Icelandic (isl_Latn):
Software	... The original text is from software data. Translate the content of json, without translating keys and placeholders (i.e. copy them rather than translate them). Produce only the Icelandic translation, without any additional explanations or commentary. Produce a valid json output that matches the input format. Please translate the following text into Icelandic (isl_Latn):

Table 4: Domain-specific system prompts. Each begins with the shared preamble “*You are a professional Icelandic translator. . . ensuring that the translation is natural.*”, elided above with ... for space.