Comparing Human and Machine Translations of Generative Language Model Evaluation Datasets

Sander Bijl de Vroe and **Georgios Stampoulidis** and **Kai Hakala** and **Aku Rouhe** and **Mark van Heeswijk** and **Jussi Karlgren**
Advanced Micro Devices, Inc.
{Sander.BijldeVroe, Georgios.Stampoulidis, Kai.Hakala, Aku.Rouhe, Mark.vanHeeswijk, Jussi.Karlgren}@amd.com

Abstract

The evaluation of Large Language Models (LLMs) is one of the crucial current challenges in the field of Natural Language Processing (NLP) and becomes even more challenging in the multilingual setting. Since the majority of the community's benchmarks exist only in English, test sets are now being machine translated at scale into dozens of languages. This work explores the feasibility of that approach, comparing a Finnish machine translation (MT) of ARC-Challenge with a new human translated version. Our findings suggest that since absolute scores are fairly close and model size rankings are preserved, machine translation is adequate in this case. Surprisingly, however, the datasets reverse the order of base models compared to their chat-finetuned counterparts.

1 Introduction and Background

Generative Large Language Models (LLMs) have made significant progress in the past few years and their usefulness is being explored in many applications. This exploration is occurring world-wide, and as such there are many multilingual models available which have been trained with data in several languages simultaneously. However, a central challenge in building multilingual models is that access to quality data in languages except for the largest ones is limited, and this challenge crucially extends to evaluation datasets.

Our ability to train acceptably performing LLMs in new languages has far outpaced our abilities to create high-quality evaluations for those languages, in part because training can rely on transfer effects, where competence acquired in one language generalises to another language to some extent (e.g., Gogoulou et al., 2021). Constructing new test sets in the language under consideration allows for controlling the quality as well as cultural validity of test items, but translating existing test sets (usually in English) to a target language involves less effort, less cost, and provides a basis to compare results across languages.

Translating entire test suites involves considerable human effort, so using automatic translation tools is an obviously attractive option. Given the immediate need to evaluate multilingual models, these machine translations of evaluation datasets have started proliferating — for example, Lai et al. (2023) automatically translate the popular benchmarks HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2020) and ARC (Clark et al., 2018) into 26 languages.

However, these strategies carry a certain risk of systematic bias and introduced error into the test, and very little work has been done to verify that the resulting evaluations can be trusted. Besides actual translation errors, sometimes the objectives of the test are rendered moot by linguistic differences: for instance, tests that exploit structural ambiguities translate poorly from an isolating language to an agglutinative one.

This work investigates how automatically translated tests compare to manually translated tests. We study the case of ARC-Challenge, the challenging subset of ARC, a four-way multiple-choice task that has become a popular English LLM benchmark. We compare the performance of several Finnish LLMs on an automatically translated version (ARC-C-fi-MT) with a new manually translated version (ARC-C-fi-HT), which we release publicly. We find that in this task setup, LLMs perform comparably on the machine- and human-translated versions, so that machine translation may actually suffice in this case. One surprising caveat is that when considering model orderings, base models outperform their chat-tuned

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 80–85

counterparts on human data, while the chat-tuned models are stronger on machine translation data.

2 Datasets and Translation Methods

2.1 ARC-Challenge

Our investigation uses versions of ARC-Challenge, the more challenging portion of the ARC dataset (Clark et al., 2018), one of the most popular evaluation datasets for LLMs. ARC is a four-way multiple-choice Question Answering (QA) dataset, drawn from grade-school science questions designed for human test takers. For example, the question "*Which of these objects is translucent?*", with choices "*A student's notebook*", "*A mirror on the bus*", "*A brick wall of the school*", "*A student's sunglass lenses*", would have the correct answer D). The Challenge portion of the corpus consists of only those questions that are answered incorrectly by an Information Retrieval (IR) system and a word co-occurrence system[1]. We translate the test split, consisting of 1172 samples.

2.2 ARC-C-fi-MT

For the machine translated version of ARC-Challenge, we use the Finnish version released by LumiOpen (2024a), also containing translations into twelve other European languages. Samples were translated using the DeepL API (DeepL, 2025a) through the DeepL Python Library (DeepL, 2025b) using default parameters.

One noteworthy limitation is that answers were translated without the context provided by the question. This carries the drawback that some answers may have an altered meaning without the context or may contain unresolvable ambiguities, although in most cases answers are long enough for correct word sense disambiguation. For example, sample `Mercury_7086520` contains the choice "*be in the same period.*", which carries a significantly different meaning in the context of the question "*Copper and gold have similar reactive properties. On the Periodic Table of the Elements, these elements are most likely to*"[2].

Example 1: Incorrect semantics
E: When making observations in nature, what is the best way for students to show respect for the environment?
F1: Miten opiskelijat voivat parhaiten huolehtia ympäristöstä ollessaan maastossa tekemässä havaintoja?
F2: Miten opiskelijat voivat parhaiten kunnioittaa ympäristöä ollessaan maastossa tekemässä havaintoja?
Example 2: Non-idiomatic translation
E: Which action would increase the amount of oxygen in a fish tank?
F1: Mikä toimi lisäisi hapen määrää akvaariossa?
F2: Mikä näistä lisäisi akvaariossa olevan hapen määrää?

Table 1: Examples: original English (E) initial inaccurate translations (*F1*) and revisions (*F2*).

2.3 ARC-C-fi-HT

Human translation data was acquired from a leading translation company that specializes in Nordic languages. The data was produced in two stages. The first version ARC-C-fi-HTv1 underwent a rigorous evaluation process by a native Finnish speaker with experience in translation and localization business. Surprisingly, a significant portion of the initial delivery was found to be of poor quality despite our guidelines. To improve translation quality, we provided detailed feedback and requested revision of the complete dataset, leading to an improved second version that we consider the gold standard, ARC-C-fi-HT.

Our feedback process focused on various difficulties, including sentence structure, semantic misinterpretations, and style. We ensured that cultural references were accurately preserved and additionally requested that a number of literal translations be corrected to more idiomatic Finnish expressions. Some indicative examples are found in Table 1. In Example 1, the first attempt F1 uses the word *huolehtia*, which translates to *take care of*. The corrected version F2 uses *kunnioittaa*, a more precise translation of *to show respect*. In Example 2, the inaccurate F1 translates *action* as *toimi*, a more formal term that usually refers to actions by organizations. The revised version F2 uses a more idiomatic phrasing (literally *which of these*), with the word for *action* omitted.

To ensure overall quality, we also established standards for capitalization, punctuation, dates, numbers, and names. The complete dataset, along

[1] As in sample VASoL_2009_5_30 mentioned above. Commonsense sentences like "A student's sunglass lenses are translucent" are unlikely in corpora, so basic strategies are less successful.

[2] DeepL chooses the translation *samalla ajanjaksolla*, a different sense of *period*.

with ARC-C-fi-HTv1 and an alternate normalized version, is available here.

3 LLM Families

We evaluate three model families, Poro, Viking and Ahma, chosen because they are effectively the only LLM families trained especially for Finnish. Note that although FinGPT (Luukkonen et al., 2023) is absent, it can be viewed as a predecessor of Poro, since Poro uses an extended version of the same training data, and the models use an identical architecture.

3.1 Poro

The Poro base model (Luukkonen et al., 2024) is a 34 billion parameter decoder-only Transformer that uses the BLOOM architecture (Le Scao et al., 2023). It was trained on 1T tokens, of which 54.5% was English, 31.7% program code, 13.0% Finnish, and 0.8% English-Finnish translation pairs.

Poro 34B Chat (Silogen, 2024; LumiOpen, 2024b) is a version of Poro 34B trained to follow instructions in both English and Finnish using full-parameter supervised finetuning. The instruction data consists of roughly 40% English, 40% Finnish, and 20% cross-lingual examples. Because such data is not readily available in Finnish, Poro 34B itself was used to translate English instruction data into Finnish.

3.2 Viking

The Viking family of models (SiloAI, 2024) is another open-source model family that covers Finnish. The models are trained on 2T tokens, which includes further Nordic languages in Danish, Icelandic, Norwegian and Swedish, along with program code. Viking uses a similar architecture as Llama 2 (Touvron et al., 2023b). In this work we experiment with the 7B and 13B variants, for which the finetuned versions are not yet released.

3.3 Ahma

The Ahma model family (Tanskanen and Toivanen, 2024) is the only family of LLMs pre-trained exclusively on Finnish data. They consist of decoder-only transformer models based on Meta's first Llama architecture (Touvron et al., 2023a). We evaluate both Ahma-7B and Ahma-3B, as well as Ahma-3B-Instruct. Note that the 7B-Instruct version is not available at the time of writing. Ahma-3B is trained for 139B tokens, while Ahma-7B was trained for 149B tokens, on a varied collection of deduplicated and detoxified Finnish text sources.

Human Translation		**Machine Translation**	
1: Poro 34B	.414	Poro 34B-C	.391
2: Poro 34B-C	.397	Poro 34B	.369
3: Viking 13B	.387	Viking 13B	.329
4: Viking 7B	.363	Ahma-7B	.327
5: Ahma-7B	.358	Viking 7B	.326
6: Ahma-3B	.324	Ahma-3B-I	.310
7: Ahma-3B-I	.323	Ahma-3B	.307

Table 2: Model rankings for ARC-C-fi-HT and ARC-C-fi-MT (acc_norm scores).

4 Methodology

We use EleutherAI's LM-evaluation-harness (Gao et al., 2024) to run the evaluations. For each of the datasets, we use the default parameters of English ARC-Challenge. In particular, we evaluate using the `multiple_choice` setting, with `doc_to_text`: "Question: {{question}}\nAnswer:" and `num_fewshot` = 0. This means that answers are obtained using logprobs instead of running inference. For each possible choice, the logprob of the choice text given `doc_to_text` is computed, and the model's answer is the maximum logprob choice.

We compute both `acc` (accuracy) and `acc_norm`. The latter metric computes accuracy when logprobs are normalized by answer length, so that longer answers are not deemed less likely only due to their length. To guarantee that results are tokenizer-agnostic, normalization is performed using number of characters rather than, for example, number of tokens.

5 Experimental Results and Analysis

One straightforward way of analyzing the translated datasets is through comparing absolute scores per model. However, the datasets can also be compared in terms of whether they preserve the ranking between two sets of evaluated models. The model rankings (Table 2) reveal two main effects. Firstly, the ordering between model sizes remains effectively constant between the different translations: clearly scores follow the expected ordering 34B $>$ 13B $>$ 7B $>$ 3B. The Ahma and

		HT	MT	EN
Poro 34B-C	acc_norm	.397	.391	.485
	acc	.376	.374	.452
Poro 34B	acc_norm	.414	.369	.462
	acc	.361	.341	.424
Viking 13B	acc_norm	.387	.329	.402
	acc	.346	.312	.359
Viking 7B	acc_norm	.363	.326	.366
	acc	.308	.301	.340
Ahma-7B	acc_norm	.358	.327	.275
	acc	.302	.276	.248
Ahma-3B-I	acc_norm	.323	.310	.250
	acc	.290	.259	.220
Ahma-3B	acc_norm	.324	.307	.255
	acc	.278	.270	.195

Table 3: Results across ARC Challenge variants with Human Translation (HT), Machine Translation (MT) and English (EN)

Viking 7B models show similar performance on Finnish — it is unlikely their change in ranking between datasets is significant.

Secondly, however, for both Poro 34B and Ahma-3B, we see a change in ordering of the base model and chat model variants (note that for other base models the finetuned versions are not yet available). The base completion models rank higher when evaluated using human translation, while the chat models rank higher for the machine translated version.

Table 3 shows full results on the three dataset versions. Here we include the accuracy results next to acc_norm for completeness. A clear initial result is that across the board for every model family, size and training method (as well as for acc and acc_norm) the absolute performance on human translated data is at least slightly higher than on machine translated data.

Still, the size of these differences varies per model. One result worth noting is that the chat models perform similarly between the two datasets in absolute terms (.397 and .391 for Poro 34B Chat; .323 and .310 for Ahma-3B-Instruct). However, and particularly for Poro, there is a larger difference for the base models (.414 and .369 for Poro 34B; .324 and .307 for Ahma 3B). Thus the chat models seem more robust across the datasets, but at a cost to performance (.369 for Poro 34B Base $<$ .391 for Poro 34B Chat on the machine translation condition).

We also find that the Poro and Viking models, trained on both English and Finnish, perform better on the English dataset than on the Finnish datasets. This is unsurprising given that in the case of Poro, there were more than four times as many English tokens in the training distribution. The Ahma models, lacking English training data, reach the expected performance of around 25% on English given 4-way multiple-choice.

6 Discussion

We propose a technique to find particularly interesting examples by 1) filtering for cases where the model is correct on one dataset and incorrect on the other and 2) sorting by the difference in logprobs on the prediction for the correct answer. In this way, we find samples where the difference in translation has the greatest effect on model prediction and performance.

This reveals some clear mistakes in machine translation. For instance, in `Mercury_SC_414274` the correct choice is *The Moon is covered with many craters*. Here the human translation is *Kuun pinnalla on paljon kraattereita* whereas machine translation outputs *Kuu on monien kraatterien peitossa*, a more literal and less fluent translation. As a result it is only on the human translation data that Poro-Chat-34B manages to select the correct answer (with a logprob of -10.2 instead of -36.0). Question `Mercury_7165218` about geology provides a more egregious error, where the choice *rift* is left untranslated as *rift*.

There are also cases, however, where flaws in machine translation actually increase model scores. In question `Mercury_SC_406710` about chameleons, the choice *hunt for food* is translated correctly by humans as *Saalistaa ruokansa*, using the verb reserved for predators, but is machine translated as *metsästää ruokaa*, using the verb for human hunting[3]. However, perhaps since *metsästää* is more common, Poro-34B-Chat correctly chooses this as the answer, whereas it fails to do so for *saalistaa*. Thus the human translation reveals the incomplete semantics of the model in this case, while the machine translation does not.

A similar case occurs in `MCAS_2014_8_6`, where Poro-34B-Chat makes the correct choice only for the machine translated answer con-

[3]Note the unresolvable ambiguity for the MT model in this case, given that it has not seen the context *chameleon*.

taining the phrase *tektonisten laattojen* (*tectonic plates*). The human translation uses the phrase *litosfäärilaattojen* (lithospheric plates), which is heavily discounted by the model at a logprob of -73.0. Both translations are correct, but *litosfäärilaattojen* is a slightly more scientific and technical term in Finnish. *tektonisten* (*of tectonics*) is perhaps more common in layman's language, which would explain both its generation by the MT system and its higher logprob in the LLM's answer. In such cases, machine translated evaluations assign an inflated accuracy to the models, which should be able to respond positively to both the common and rarer scientific terms. For an agglutinative case-based language such as Finnish, similar cases would be possible when a human translator chooses a more accurate but less common grammatical case.

There are many future research avenues here. One option is to further investigate the chat and completion model reordering. This is possibly explained by an alignment of the fine-tuning training data with machine translation data — in both cases, models are trained using curated sentence pairs (whereas base model pre-training data consists of large chunks of text from massive corpora that tend to be less curated). Perhaps, then, fine-tuning a base model pulls it in the direction of the machine translation model distribution. Future work that compares more pairs of base and chat models, along with extended logprob analyses of both models types, may elucidate the picture.

Future work will also investigate the complex set of benefits and drawbacks of human translation. Human subjectivity and inconsistencies in judgment may introduce bias, and from a practical standpoint manual reviews can be time-consuming and expensive. One concrete direction is to compare the gold standard to ARC-C-fi-HTv1 and versions with alternative choices of normalization. It would also be worthwhile to explore alternate MT solutions, especially ones in which the models have access to the question as context when translating the answers.

7 Conclusion

Following the recent trend to machine translate English evaluation datasets at scale, this work compares a new human translation of ARC-Challenge into Finnish with a machine translated version. Our results indicate that for Finnish ARC-Challenge, the machine translated dataset rivals the usefulness of the HT dataset for comparative evaluation of LLMs.

This is observed through the small absolute differences between scores (with models performing slightly more favorably on human translations as expected), as well as through the preservation of ordering of model sizes. One interesting caveat is that while chat-finetuned models outperform base models on machine-translated evaluation data, base models actually outperform their chat-finetuned counterparts on the human translated data, warranting further investigation.

Thus although there are drawbacks to using machine translation, especially for literature or other longer-form data, this work reveals that for comparative evaluation of Finnish language models on short multiple-choice questions, MT is sufficient. Future work can continue to reveal distributions of evaluation data, language translation pairs and model classes where this holds. It is clear that the intersection of translation and LLM evaluation provides unique challenges and opportunities that now deserve more attention than ever.

Acknowledgments

We thank Jonathan Burdge, Elaine Zosa, and Teppo Lindberg for their helpful comments. We also thank Sanna Piha for valuable contributions in dataset acquisition and review, and the AMD technical reviewers and copy-editors for their insights and advice. This work has been supported by the European Commission through the DeployAI project (grant number 101146490). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or DeployAI. Neither the European Union nor DeployAI can be held responsible for them. AMD is a trademark of Advanced Micro Devices, Inc. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies.

References

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge.

DeepL. 2025a. Deepl. `https://github.com/`

DeepLcom/deepl-python. Accessed 2nd January 2025.

DeepL. 2025b. Deepl python library. https://www.deepl.com/en/translator. Accessed 2nd January 2025.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Evangelia Gogoulou, Ariel Ekgren, Tim Isbister, and Magnus Sahlgren. 2021. Cross-lingual transfer of monolingual models. *arXiv preprint arXiv:2109.07348*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

LumiOpen. 2024a. Lumiopen/ arc_challenge_mt. Accessed on September 10th 2024.

LumiOpen. 2024b. Poro-34b-chat. Accessed on May 28th 2024.

Risto Luukkonen, Jonathan Burdge, Elaine Zosa, Aarne Talman, Ville Komulainen, Väinö Hatanpää, Peter Sarlin, and Sampo Pyysalo. 2024. Poro 34b and the blessing of multilinguality.

Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, Thomas Wang, Nouamane Tazi, Teven Scao, Thomas Wolf, Osma Suominen, Samuli Sairanen, Mikko Merioksa, Jyrki Heinonen, Aija Vahtola, Samuel Antao, and Sampo Pyysalo. 2023. FinGPT: Large generative models for a small language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2710–2726, Singapore. Association for Computational Linguistics.

SiloAI. 2024. Viking 7b: The first open llm for the nordic languages. Accessed on October 7th 2024.

Silogen. 2024. Poro 34b chat is here. Accessed on May 28th 2024.

Aapo Tanskanen and Rasmus Toivanen. 2024. Ahma-3b (revision 0b51e96).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.