Benchmarking Frontier LLM Understanding in Latvian

Arturs Kanepajs*

October 16, 2024

Abstract

This study evaluates the performance of large language models (LLMs) on the MMLU (Massive Multitask Language Understanding) benchmark in Latvian versus English, assessing the language gap and determining the best-performing model in Latvian. Using 112 translated MMLU questions, six models were tested on Latvian (machine-translated and human-redacted) and English versions. Results show a variable English-Latvian performance gap across models, with the best performing models overall also exhibiting minimal language gaps.

1 Introduction

The potential benefits of advanced AI are vast, but to ensure these advantages are globally accessible, it's crucial that AI systems perform well across multiple languages. Previous research has highlighted a significant disparity between the performance of frontier large language models in English compared to other languages, particularly those with limited resources [3, 6, 8, 4].

As AI models continue to evolve, it's essential to monitor how this language gap is narrowing. Users working with models in various languages could greatly benefit from comparative performance analyses across different linguistic contexts. However, evaluating model performance in non-English languages presents challenges. While automated translation of benchmarks is cost-effective, it raises concerns about quality. Conversely, human translations, though potentially more accurate, can be prohibitively expensive.

Driven by these considerations and leveraging the author's fluency in Latvian, this study aims to address the following key questions:

- Which frontier large language model demonstrates the highest proficiency in Latvian?
- What is the extent of the performance gap between English and Latvian?
- To what degree is human involvement necessary in the translation of benchmarks?

I utilize the MMLU (Massive Multitask Language Understanding) benchmark, which covers 57 subjects ranging from STEM to humanities and social sciences.

2 Methodology

1. The MMLU dataset [5] A subset of 112 questions with answers was translated into Latvian using the MyMemory API [7]
2. A Latvian-fluent annotator (the author) reviewed and edited the translations, focusing on correcting errors that could hinder question comprehension or lead to misinterpreted answer options.

*akanepajs@gmail.com

3. Six large language models were evaluated for performance.[1] UK AISI Inspect framework [1] was used as a basis for writing the evals. served as the foundation for developing the evaluation scripts. All datasets and code are accessible on GitHub.[2]

The models' performance was assessed with the following datasets:

A The original 112 questions in English

B The unedited machine translation of these questions in Latvian

C The human-edited translation of these questions in Latvian

3 Results

The performance of the models across different languages and conditions is presented in Table 1).

Model	Latvian (autotranslated & redacted)	English	Latvian (autotranslated)
o1-preview	0.848 [0.783, 0.913]	0.875 [0.814, 0.936]	0.821 [0.752, 0.890]
Gemini 1.5 Pro	0.786 [0.713, 0.859]	0.846 [0.780, 0.912]	0.732 [0.653, 0.811]
ChatGPT 4o	0.759 [0.683, 0.835]	0.821 [0.752, 0.890]	0.723 [0.643, 0.803]
Claude 3.5 Sonnet	0.756 [0.679, 0.833]	0.857 [0.793, 0.921]	0.714 [0.633, 0.795]
Llama 3.1 405B	0.688 [0.605, 0.771]	0.839 [0.772, 0.906]	0.643 [0.556, 0.730]
Mistral Large 2	0.580 [0.490, 0.670]	0.768 [0.693, 0.843]	0.580 [0.490, 0.670]
AVERAGE	**0.736 [0.703, 0.769]**	**0.834 [0.807, 0.861]**	**0.702 [0.668, 0.736]**

Table 1: Comparison of model performance across different languages and conditions, with 95% confidence intervals.

3.1 Which model understands Latvian best?

OpenAI's o1 model achieves the highest performance in Latvian, although the differences between o1, ChatGPT 4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro are not statistically significant. In contrast, Llama 3.1 405B and Mistral Large 2 demonstrate significantly lower performance (see Figure 1).

3.2 How big is the English-Latvian performance gap?

The average performance gap between English and Latvian, calculated by pooling observations, is 9.8% (95% c.i: 5.4% to 14.2%). This difference is comparable to approximately two-thirds of the performance gap between GPT-3.5 and GPT-4 in English [9] owever, the gap narrows for the top-performing models and is only statistically significant for Mistral Large 2 (see Figure 2).

3.3 How important is human involvement

Model performance with human-redacted translations shows a modest improvement (3.4% on average, see Table 1) ompared to unredacted machine translations. However, this difference is not statistically significant (95% c.i: -1.2% to 8.2%). This finding may raise questions about the necessity of human involvement in such evaluations, at least for Latvian. It's worth noting that a free translation service was used; state-of-the-art machine translation might further reduce the benefits of human involvement.

4 Discussion

- Anecdotal evidence showed that some of the tested models were much better at translating questions and answers than the free translation service. Future research could make use of

[1] o1-preview-2024-09-12 , gpt-4o-2024-08-06, claude-3-5-sonnet-20240620, gemini-1.5-pro-002, Meta-Llama-3.1-405B-Instruct-Turbo, and Mistral Large 2 (obtained for API by calling "mistral-large-latest". Temperature=0.5 was used for all models except for o1-preview, for which only temperature=1 setting was available.

[2] `https://github.com/akanepajs/capabilities_lv`

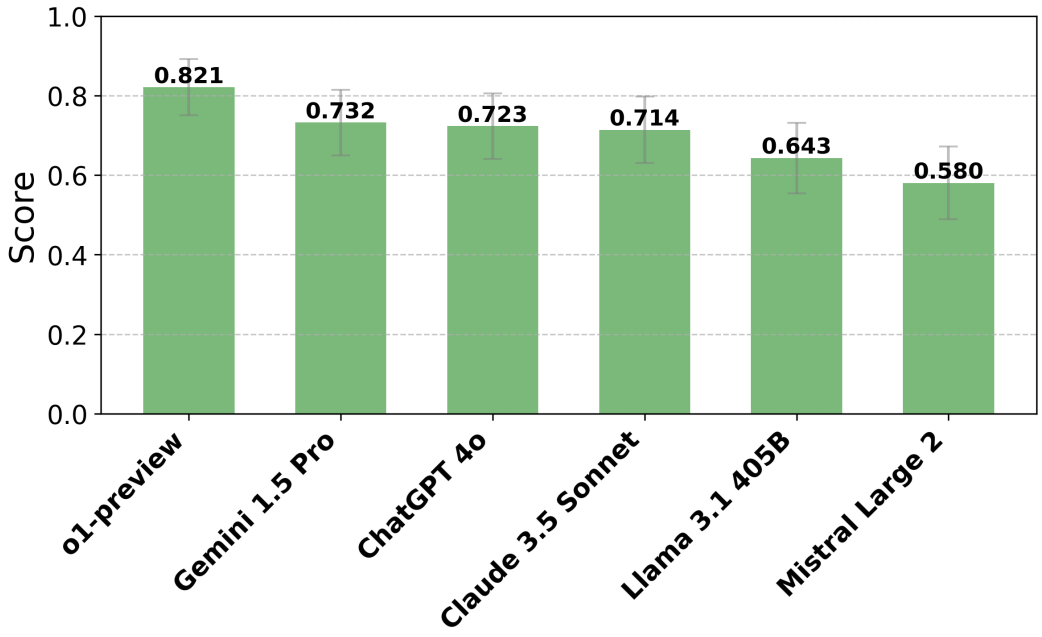

Figure 1: **Model performance in Latvian. Randomly selected MMLU benchmark questions, autotranslated from English & redacted, n=112.**

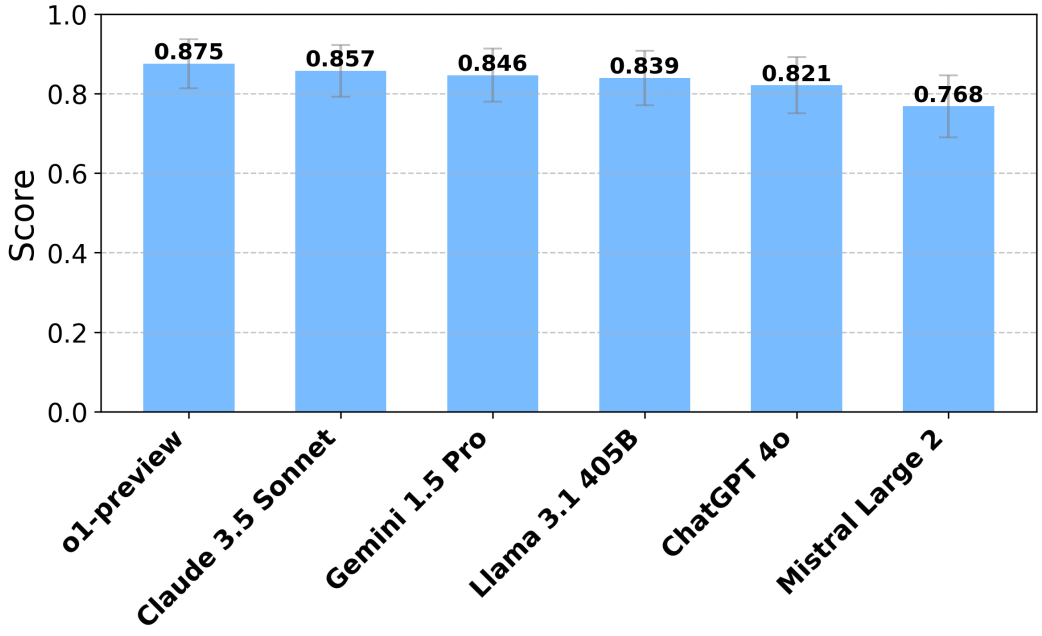

Figure 2: **Model performance in English. Randomly selected MMLU benchmark questions. n=112.**

the LLM translation capabilities. However, it is important not to bias the results in favor of one model or another: it is not inconceivable that a given model finds its own translations easier to interpret than those of other models (which is another hypothesis to explore). Alternatively, it is possible to use other translation services and human translation services together or separately.

- These, as other benchmark results, may be subject to bias due to potential data contamination. [2]. The English MMLU dataset is more likely to have been included in or influenced the models' training data. This could lead to an overestimation of the performance gap between languages, as models might have prior exposure to the English questions.
- Cultural context introduces another potential source of bias and reduced relevance in this study. For example, Professional Law questions are based on the U.S. legal system, not Latvian law, which may confuse models when presented with questions in Latvian. Future research could assess the impact of cultural context by using a larger sample size and analyzing model performance in culturally sensitive subcategories like Professional Law. However, U.S.-centric legal questions are inherently limited in evaluating legal expertise within the Latvian context. Adapting such questions to local contexts is crucial but may require costly specialist knowledge.
- Expanding the sample size in future studies could yield more robust results. The scope of this investigation was primarily constrained by the human resources required for translation redaction and the available computational power.

5 Conclusion

The research reveals that while an English-Latvian performance gap exists in large language models, it varies significantly among them. Top-performing models show minimal differences between languages. Human redaction of machine-translated questions had little impact, suggesting automated translations might suffice for Latvian evaluations. This study contributes to understanding AI capabilities in non-English languages and identifies areas for future research, including potential biases and leveraging models' translation abilities for benchmark creation.

6 Acknowledgements

I am grateful to Jonas Kgomo for advice on how to conduct evaluations.

References

[1] AI Safety Institute. Inspect - ai safety institute. https://inspect.ai-safety-institute.org.uk/, 2024. Accessed: 2024-10-15.

[2] Andrew M. Bean, Simi Hellsten, Harry Mayne, Jabez Magomere, Ethan A. Chi, Ryan Chi, Scott A. Hale, and Hannah Rose Kirk. LINGOLY: A Benchmark of Olympiad-Level Linguistic Reasoning Puzzles in Low-Resource and Extinct Languages, 2024. _eprint: 2406.06196.

[3] Cohere For AI team. Policy Primer - The AI Language Gap, 2024.

[4] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and others. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024.

[5] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.

[6] Artūrs Kanepajs, Vladimir Ivanov, and Richard Moulange. Towards safe multilingual frontier ai, 2024.

[7] MyMemory. Mymemory translation memory - api documentation. https://mymemory.translated.net/doc/spec.php, 2024. Accessed: 2024-10-15.

[8] OpenAI. O1 system card. Technical report, OpenAI, September 2024. Accessed on October 16, 2024.

[9] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather

Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 Technical Report, 2024. _eprint: 2303.08774.

A Translation Examples

This appendix provides examples of MMLU questions in their original English form, followed by their Latvian machine translations, and finally the human-redacted Latvian versions. These examples illustrate the translation and redaction process used in this study.

Question	A	B	C	D	Ans
Which of the following items below is NOT considered a barrier to the diffusion of an innovation?	Internet	Distance	Cultural obstacles	Government interference	A
When we convert total number of errors on a test to percent correct, or vice versa, we are performing which type of transformation	Linear	Nonlinear	Curvilinear	Cumulative	A
Both over-the-counter niacin and the prescription drug Lipitor are known to lower blood cholesterol levels. In one double-blind study Lipitor outperformed niacin. The 95% confidence interval estimate of the difference in mean cholesterol level lowering was (18, 41). Which of the following is a reasonable conclusion?	Niacin lowers cholesterol an average of 18 points, while Lipitor lowers cholesterol an average of 41 points.	There is a 0.95 probability that Lipitor will outperform niacin in lowering the cholesterol level of any given individual.	There is a 0.95 probability that Lipitor will outperform niacin by at least 23 points in lowering the cholesterol level of any given individual.	None of the above.	D
Noam Chomsky and B. F. Skinner disagreed about how children acquire language. Which of the following concepts is most relevant to the differences between their theories?	phonemes	morphemes	linguistic relativity hypothesis	language acquisition device	D
According to Huemer, even if drug use harms a person's friends, families, and other relations,	this would still not justify drug prohibition.	this would justify drug prohibition.	this would only justify drug prohibition if drug use was extremely likely to cause these harms.	this would only justify drug prohibition if drug use was more likely to cause these kinds of harms than other prohibited activity.	A

Table 2: MMLU questions in English

Question	A	B	C	D	Ans
Kurš no šiem elementiem NETIEK uzskatīts par šķērsli inovāciju izplatīšanai?	Internets	attālums	Kultūras šķēršļi	Valdības iejaukšanās	A
Konvertējot kopējo kļūdu skaitu testā uz procentuālo pareizību vai otrādi, mēs veicam kāda veida transformāciju	Lineāraquadratic filter mode	Nelineārs	Līklīnija	Uzkrāts	A
Ir zināms, ka gan bezrecepšu niacīns, gan recepšu zāles Lipitor pazemina holesterīna līmeni asinīs. Vienā dubultmaskētā pētījumā Lipitor pārspēja niacīnu. Vidējā holesterīna līmeņa pazemināšanas atšķirības 95% ticamības intervāla novērtējums bija (18, 41). Kurš no šiem secinājumiem ir pamatots?	Niacīns pazemina holesterīna līmeni vidēji par 18 punktiem, bet Lipitor pazemina holesterīna līmeni vidēji par 41 punktu.	Pastāv 0,95 varbūtība, ka Lipitor pārspēs niacīnu jebkura indivīda holesterīna līmeņa pazemināšanā.	Pastāv 0,95 varbūtība, ka Lipitor pārspēs niacīnu vismaz par 23 punktiem, samazinot jebkura indivīda holesterīna līmeni.	Nevienā no minētajām	D
Noams Čomskis un B. F. Skiners nepiekrita tam, kā bērni apgūst valodu. Kurš no šiem jēdzieniem ir visatbilstošākais atšķirībām starp to teorijām?	fonēmas	morfēmas	lingvistiskās relativitātes hipotēze	valodas apguves ierīce	D
Saskaņā ar Huemer teikto, pat ja narkotiku lietošana kaitē cilvēka draugiem, ģimenēm un citām attiecībām,	tas joprojām neattaisnotu narkotiku aizliegumu.	tas attaisnotu narkotiku aizliegumu.	tas attaisnotu narkotiku aizliegumu tikai tad, ja narkotiku lietošana varētu radīt šo kaitējumu.	tas attaisnotu narkotiku aizliegumu tikai tad, ja narkotiku lietošana varētu radīt šāda veida kaitējumu vairāk nekā citas aizliegtas darbības.	A

Table 3: MMLU questions autotranslated to Latvian

Question	A	B	C	D	Ans
Kurš no šiem elementiem NETIEK uzskatīts par šķērsli inovāciju izplatīšanai?	Internets	attālums	Kultūras šķēršļi	Valdības iejaukšanās	A
Konvertējot kopējo kļūdu skaitu testā uz procentuālo pareizību vai otrādi, mēs veicam kāda veida transformāciju	Lineāru	Nelineāru	Līklīniju	Uzkrāto	A
Ir zināms, ka gan bezrecepšu niacīns, gan recepšu zāles Lipitor pazemina holesterīna līmeni asinīs. Vienā dubultmaskētā pētījumā Lipitor pārspēja niacīnu. Vidējā holesterīna līmeņa pazemināšanas atšķirības 95% ticamības intervāla novērtējums bija (18, 41). Kurš no šiem secinājumiem ir pamatots?	Niacīns pazemina holesterīna līmeni vidēji par 18 punktiem, bet Lipitor pazemina holesterīna līmeni vidēji par 41 punktu.	Pastāv 0,95 varbūtība, ka Lipitor pārspēs niacīnu jebkura indivīda holesterīna līmeņa pazemināšanā.	Pastāv 0,95 varbūtība, ka Lipitor pārspēs niacīnu vismaz par 23 punktiem, samazinot jebkura indivīda holesterīna līmeni.	Neviens no minētajām	D
Noams Čomskis un B. F. Skiners nebija vienisprātis par to, kā bērni apgūst valodu. Kurš no šiem jēdzieniem ir visatbilstošākais atšķirībām starp viņu teorijām?	fonēmas	morfēmas	lingvistiskās relativitātes hipotēze	valodas apguves ierīce	D
Saskaņā ar Huemer teikto, pat ja narkotiku lietošana kaitē cilvēka draugiem, ģimenēm un citām attiecībām,	tas joprojām neattaisnotu narkotiku aizliegumu.	tas attaisnotu narkotiku aizliegumu.	tas attaisnotu narkotiku aizliegumu tikai tad, ja narkotiku lietošana varētu radīt šo kaitējumu.	tas attaisnotu narkotiku aizliegumu tikai tad, ja narkotiku lietošana varētu radīt šāda veida kaitējumu vairāk nekā citas aizliegtas darbības.	A

Table 4: MMLU questions autotranslated to Latvian, and human-redacted