"I Need More Context and an English Translation": Analysing How LLMs identify Personal Information in Komi, Polish, and English

Nikolai Ilinykh
CLASP, FLoV
University of Gothenburg, Sweden
nikolai.ilinykh@gu.se

Maria Irena Szawerna
Språkbanken Text, SFS
University of Gothenburg, Sweden
maria.szawerna@gu.se

Abstract

In this paper we present a pilot study and a qualitative analysis of the errors made by three large language models (LLMs) prompted to identify personal information (PI) in texts written in languages with varying resource availability: Komi (extremely low), Polish (medium), and English (high). Our analysis shows that LLMs perform better in detection of PI when provided with JSON-eliciting prompts. We also conjecture that the rich morphology and inflectionality of languages like Komi and Polish might affect the models' performance. The small-scale parallel dataset of text that we introduce here can be used as a starting point in developing benchmarks for evaluation of PI detection with longer textual contexts and LLMs.

1 Introduction

The lack of data for *low-resourced* languages is a known problem in computational linguistics. This problem can result in biases "within and across societies" (Søgaard, 2022), since the speakers of such languages can effectively be excluded from using language technology tools. Building infrastructure that uses such technology as LLMs to study and preserve low-resourced languages is important.

A key concern in the development of NLP infrastructure is the privacy of the data subjects and other individuals mentioned.[1] Linguistic data typically includes names, family relationships, health status, or other sensitive details, especially when collected texts are personal conversations, narratives, or interviews (Szawerna et al., 2024), and even seemingly scarce or incomplete PI may be used to reidentify the data subject (Salehi et al., 2017) and result in discrimination based on, for example, medical conditions or faith. Methods to obfuscate identities of data subjects have long been employed in linguistics (Thomas, 2010; Wang et al., 2024), but only a few of them have used computational approaches for identification of PI in low-resourced languages like Komi (Hämäläinen et al., 2023). Since personal information has been found in data used to train LLMs for many high-resourced languages – raising concerns about potential leaks in their outputs (Subramani et al., 2023) – it is crucial to protect privacy of data in these languages. However, protecting personal information in low-resourced languages is especially important as these languages already struggle with limited datasets, funding, and institutional support, making them particularly vulnerable to privacy risks.

In this pilot study we take a step towards better PI detection in low-resourced languages and prompt currently available LLMs. Such models, trained on multilingual corpora, can be prompted to perform a range of tasks, from text classification to text generation, even in languages where limited training data is available (A Pirinen, 2024; Purason et al., 2024b). LLMs have been studied in the context of low-resourced Uralic languages for the task of POS tagging (Alnajjar et al., 2024). They have also been used to support the creation of online dictionary tools (Alnajjar et al., 2020). The role of LLMs in PI detection in high-resourced language like English and Chinese has started being explored only recently (Yang et al., 2023), while their role in the context of low-resourced languages for PI detection remains unexplored.

To facilitate research in that direction, here we analyze the differences in the behavior of three LLMs in PI detection in languages with varied resource availability and linguistic structure. We use text data from two Uralic languages, Komi-Permyak and Komi-Zyrian. We construct a paral-

[1] For more on legal requirements regarding privacy in EU, see Official Journal of the European Union (2016).

Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL 2025), pages 165–178

lel corpus containing Komi sentences[2] with their Polish and English translations. We prompt Llama 3.1 with 8B parameters (Grattafiori et al., 2024), Mistral 7B (Jiang et al., 2023), and Gemma 2 with 9B parameters (Gemma Team et al., 2024) for PI detection and test six different prompt configurations. Our contributions, therefore, consist of 1) **a small, native speaker-curated parallel corpus of sentences** containing potential personal information in Komi, English, and Polish[3], and 2) **an initial analysis** of how three LLMs perform on the aforementioned dataset with respect to language's resource availability and inflectionality.

2 Materials and Methods

Data We looked at the Universal Dependencies treebanks for Komi-Permyak and Komi-Zyrian (Rueter et al., 2020; Partanen et al., 2018; Zeman et al., 2024) and found that there are 366 sentences in which there is at least one word that is labeled with one of the semantic tags for proper nouns as used in the GiellaLT infrastructure (Pirinen et al., 2023). These semantic tags classify names and nouns into categories such as animal (Sem/Ani), female (Sem/Fem) and male names (Sem/Mal), objects (Sem/Obj), organisations (Sem/Org), places (Sem/Plc), surnames (Sem/Sur), and web addresses (Sem/Web). Blokland et al. (2020) have previously used these semantic tags to identify nouns which are possible instances of PI in a rule-based PI detection system.

Among the sentences with semantic tags for proper nouns, 170 were translated to English and Polish by authors of this study. The sentences were first translated by the first author of this study (a native Komi-Permyak speaker and a proficient English speaker) from Komi-Permyak and Komi-Zyrian to English with the help of Neurotõlge[4] (Yankovskaya et al., 2023; Purason et al., 2024a), Google Translate[5] and Majbyr Translate[6]. Polish translations were created by the second author (a native Polish speaker and a proficient English speaker) based off of the English translations, and with the help of Google Translate in some cases. The original names of people and places were preserved during translation into English and the final form of the translated sentence was always overseen by a human. In the end, our data included 35 sentences with female names, 47 sentences with male names, 49 sentences with place names, and 39 sentences with surnames in them. Some sentences contain more than one name, possibly of different types. Importantly, more information that could be considered personal and which does not necessarily belong to the aforementioned types may be found in sentences, and was impossible to account for during the sentence extraction process. Our resulting dataset can be accessed on Zenodo via https://zenodo.org/records/14845329.

PI categories	Text	JSON
PI only	Prompt 1	Prompt 2
Megyesi et al. (2018)	Prompt 3	Prompt 4
Subramani et al. (2023)	Prompt 5	Prompt 6

Table 1: Prompts by tag and output type.

Models and prompts We tested three multilingual pre-trained large language models: Llama 3.1 with 8B parameters (Grattafiori et al., 2024), Mistral 7B (Jiang et al., 2023), and Gemma 2 with 9B parameters (Gemma Team et al., 2024)[7]. The models and their weights were accessed via Ollama[8]. Uploading data containing PI to third-party services is not optimal, which is why we chose models that we were able to run locally. Note that we chose recent LLMs which are similar in size and comparable.

We used six different one-shot prompts, passed to the models together with the sentences, following the official guide on prompting Llama models[9], with a similar structure to the one used by Yang et al. (2023) for PI detection. The prompts varied in terms of (i) the output format (produce a sentence with PI instances replaced with appropriate tags or a JSON structure) and (ii) the PI classification.

[2] Originating from Komi corpora (Rueter et al., 2020; Partanen et al., 2018; Zeman et al., 2024); we feature 143 sentences in Komi-Zyrian and 27 sentences in Komi-Permyak.

[3] It is important to highlight that there exists no comprehensive definition of what it means to be a *low-resourced* language (Nigatu et al., 2024); traditionally, due to small amounts of available data among other things, Komi and many other Uralic languages have been considered low-resourced. Polish boasts a significantly larger collection of corpora, tools and models than Komi (Dadas, 2019), and has been positioned as the higher-resourced counterpart of West Slavic minority languages such as Kashubian, Silesian, or Sorbian (Torge et al., 2023; Rybak, 2024), but in comparison with English, its resources are still very limited.

[4] https://translate.ut.ee

[5] https://translate.google.com

[6] https://translate.majbyr.com

[7] In the paper we refer to these models as **Llama**, **Mistral**, and **Gemma** respectively.

[8] http://ollama.com

[9] https://www.llama.com/docs/how-to-guides/prompting/

System: You are a multilingual personal information detection tool. Personal information is information that can lead to someone in the text being reidentified, like their name, surname, middle name, patronymic, nickname, where they live, address, city, country, zip code, where they work, study, or spend a lot of their time, what unique lines or modes of transport they travel with, their age, any dates mentioned in the text, phone numbers, personal identity numbers, bank account numbers, other number sequences, e-mail addresses, urls, their work titles, education, types of family relations, information about faith, political beliefs, sexuality, ethnicity, unique achievements, etc.

User: For each token in the given text, determine whether it is a piece of personal information. Return the text with "PI" replacing every instance of personal information.

Example:
Text: I'm from Slovakia , but one of my best friends , Marie , is from Norway .
Result: I'm from PI , but one of my best friends , PI , is from PI.

Text: [PLACEHOLDER]
Result:

Figure 1: One of the prompt templates used in this study. When fed to a model, [PLACEHOLDER] is replaced with an actual text.

For PI classification we used different category formulations: (1) a "PI" category encompassing all personal information, (2) detailed name- and geographical location-related categories inspired by Megyesi et al. (2018), and (3) a slightly re-phrased PI categorization from Subramani et al. (2023). See Table 1 for a summary of the combinations. All of the prompts included the description of the task, tags, output format, and a single example of an input-output pair followed by the input that the model should generate output for. Examples can be found in Figure 1 and Appendix A.

3 General error analysis

After feeding the models the prompt–sentence combinations, we collected their outputs, which we subsequently manually analyzed. We begin with an analysis of the errors encountered and then proceed to examine two specific examples. In this pilot study we did not run any quantitative analysis, as the data we have lacks token-level annotation of PI in two of the three languages. Moreover, the annotation that we do have for Komi is using the GiellaLT tags, and not the aforementioned categories (1-3); thus, our analysis is preliminary.

Komi **Gemma** ignores case markers in words identified as PI. For example, in the Komi-Zyrian sentence *Сійӧ быдмис Парижын, Францияса юркарын* 'He grew up in Paris.INE[10], the capital.INE of France.LOC', the model marks only part of the word *Францияса* 'France.LOC', ignoring the marker *-са*. In contrast, it marks the entire word when it appears in the genitive case in Komi-Permyak, e.g. *Франциялӧн* 'of France.GEN'. When asked to tag PI, **Gemma** often misidentifies the language as Russian or Urdmut and translates the text into English. It also frequently asks for more context to identify PI, refusing to produce an output. **Llama** rarely provides output, referring to concerns about revealing information that could lead to reidentification. Models are good at identifying first names and surnames (albeit worse with culture-specific names), but they struggle with names of places. For example, **Gemma** mistakenly tags *Парижын* 'in Paris.INE' in both Komi varieties as `situation` and *Франция* 'France.NOM' as `society`. **Llama** detects spans correctly, but often assigns the wrong tag: labelling *быдмис* 'grew up' in both Komi varieties as `birth`, *Парижын* 'in Paris.INE' as `records`, *Франция* 'France.NOM' as `birth` and *юркарын* 'the capital.INE' as `society`. While **Mistral** provides output in the requested format, it struggles with tagging, changes spelling and produces many hallucinations. For example, it marks the personal pronoun *Сійӧ* 'he/she.NOM' in Komi-Zyrian as `PI` and completely alters the initial sentence from *Сійӧ быдмис Парижын, Францияса юркарын* 'He grew up in Paris.INE, the capital.INE of France.LOC' to *PI абыдмис Пиржин, PI юркарын*, where only *юркарын* 'the capital.INE' is a correct word.

English **Gemma** can not only mark a name as PI but also sometimes identify and tag related pronouns when asked to provide output in JSON format, e.g. *[...] replied **Galina**, with a dry smile from the corner of **her** mouth [...].* However, it does not always follow the instructions and sometimes invents tags that are not part of the tagset, such as `<other>` for ambiguous PI categories. In one instance it is also able to assign the `<social>` tag to *Comrade* and `<character>` to *Voroshilov*, where the latter is a surname and the former is a noun referring to *Voroshilov*. **Llama** generates extensive explanations and often refuses

[10] Morphological analysis for Komi words was conducted with the help of uralicNLP: https://github.com/mikahama/uralicNLP?tab=readme-ov-file

to perform the task, mirroring its behavior on Komi. It also hallucinates tags and fails to mask multi-token PI spans accurately such as tagging only *Voroshilov* as `<firstname_male>` in *Comrade Voroshilov*. While deciding whether *Comrade* is a part of a PI span can be problematic, it is not unprecedented to find such titles included in the span: Pilán et al. (2022) include elements like *Mr.* or *Dr.* into the same span as the name and surname. Therefore, it is possible that inclusion of *Comrade* in reference to *Voroshilov* can lead to reidentifcation of this person in a different situation. **Mistral**, while hallucinating and omitting many PI instances, performs better at masking anglophone names. For example, it correctly masks names like *Mary*, *Peter*, and *Jane* using appropriate tags. However, it fails to mask names such as *Svezhov* (ko.: *Свежов)*, *Petya* (ko.: *Петя)*, or *Sasha* (ko.: *Саша)*. Additionally, it masks *Masha* (ko.: *Маша)* as `<firstname_unknown>`, indicating a lack of understanding of the name's gender (typically female). All models demonstrate (i) a tendency to over-generate and provide unrequested explanations that are difficult to evaluate and (ii) struggle with maintaining consistency in tag assignment.

Polish **Gemma** appears to misclassify inflectional cases of the words thus assigning it to the wrong gender. For example, in the sentence *[...] tuż obok domu Epimowa Punegowa* '[...] in the immediate vicinity of the Epimov.GEN Punegov.GEN house' the model mistakenly assigns *Epimowa* and *Punegowa* to `<surname_female>` and `<surname_male>` respectively, while both these are male names. **Llama** refuses to perform the task stating that it cannot give away information that could lead to someone being re-identified. It also incorrectly identifies same words in same sentences across multiple prompts: under two different prompts it tags *cerata* 'oilcloth' as either a street name or a type of a document. **Mistral**'s output is not supplemented by extensive explanations, but the model tends to hallucinate and produce incorrect tags. For example, when asked to mark personal information as `PI` in *Dorósł w Paryżu, stolicy Francji* 'He grew up in Paris.INS, the capital of France.GEN', while the model assigns `PI_City` to *Paryżu* and `PI_Country` to *Francji*, it also incorrectly assigns `PI_Name` to *Dorósł* which is a verb. Note that these tags are hallucinated - they are not like the ones we have prompted the model to produce. **Mistral** also often translates Polish sentences into English in its output.

3.1 Case analysis

We analyze outputs produced by **Gemma** for the prompt 4 as specified in Table 1, because **Gemma** has shown to be the most consistent in the quality of its outputs. Each example has output for English (top) and tokenized output for Komi-Permyak (middle) and Polish (bottom). We will focus on the two characters mentioned in each sentence: *Petya* and *Masha*. The main reason for choosing these sentences in particular for comparison is that at a first glance, they only differ in terms of what verb they feature. However, in Komi and Polish, these two verbs have a different influence, eliciting specific case endings in the object of the sentence (*Masha*). By comparing these two sentences we can therefore investigate how the model handles this grammatical and morphological diversity.

(1)
F-M			*F-F*	
Petya	*befriends*		*Masha*	.
F-U			S-U	
Петя	ёртасьӧ		Машаӧт	.
F-U			F-F	
Petja	zaprzyjaźnia się z		Maszą	.

(2)
F-M		*F-F*				
Petya	*loves*	*Masha*				.
F-U		S-U				
Петя	любит	Маша	о	с		.
F-M		S-F				
Petja	kocha	Maszę				.

In both of the examples in English the model correctly assigned `<firstname_male>` to *Petya* and `<firstname_female>` to *Masha*, suggesting that the semantic difference in the verbs has no effect between these two sentences, at least in English. In example 1, in Komi-Permyak, the model marked Петя 'Petya.NOM' as `<firstname_unknown>` and Машаӧт 'Masha.COM' as `<surname_unknown>`. For Polish, it identified *Petja* 'Petya.NOM' as `<firstname_unknown>` and *Maszą* 'Masha.INS' as `<firstname_female>`. While *Petya* is marked correctly as `<firstname_male>` when given English text, the model cannot identify the gender in Komi-Permyak and Polish. The model also thinks that Машаӧт 'Masha.COM' is a surname without gender indicator. Mistakes like this (the model thinks there is e.g. no gender indicator) might result in leakage of situational and societal

context, because the affix *-кöт* in *Машакöт* indicates comitative case that is used to express companionship, and this type of information can be considered personal. In example 2, the model seems to now identify *Petja* 'Petya.NOM' in Polish as `<firstname_male>`, while thinking that *Maszę* 'Masha.ACC' is an instance of `<surname_female>`. For Komi-Permyak, the model translates the example into Russian (the original text is *Петя любитö Машаöс* 'Petya.NOM loves Masha.ACC') and tokenizes the affix. This example demonstrates a fragile behavior of **Gemma** and combined with our general error analysis, suggests that models often try to translate input in less familiar language to a language that is more known to them (English, Russian). While Russian and Komi-Permyak share the cyrillic alphabet, the similarities and grammatical differences between two languages cannot be exploited by LLMs, because intricacies in less-resourced languages are then reduced to phenomena in a language with more resources.

4 Discussion and conclusions

Our small qualitative examination suggests that across the languages, models, and prompts that we tested, **Gemma** with JSON-eliciting prompts performs best. Overall, the models exhibited the best performance on the English sentences, followed by Polish, with Komi being the most difficult. One problem for the models is the rich morphology of Komi variants and Polish. The models also denote that they lack context to make a judgment, which highlights the difficulty in disambiguating whether a piece of information is personal or not. They are often trying to default to English or ask for an English translation when asked to perform the task on a low-resourced language that they cannot recognize. The models also learn differently from various tagsets: the one from Subramani et al. (2023) is hard to generalize from, while tags based on Megyesi et al. (2018) appear to be assigned correctly more often. Non-anglophone names, especially Komi ones, are hard for models to tag, especially in terms of the gender.

While Yang et al. (2023) consider their findings for high-resourced languages to be promising, we consider it better to err on the side of caution regarding any conclusions on the performance of LLMs on PI identification task for low-resourced languages. Our impression is that even though the models' perform well on other NLP tasks, in this case, their outputs require manual post-processing and are not immune to hallucinations. This makes them highly unreliable for the incredibly high-stakes task of PI detection on their own with the prompts used, even for high-resourced languages, but especially for the low-resourced ones, where the error rate appears to be higher. Future work should focus on evaluating models on longer texts with more context and further refinement of the best-performing prompts.

This is — to the best of our knowledge — the first study investigating the performance of LLMs on PI detection in more than just high-resourced languages (specifically, in such low-resourced language as Komi), and the first one examining how LLMs handle inflectionality in this task. We are also contributing a novel parallel dataset translated by native speakers. We hope that our work will inspire more research on the topics within the intersection of LLMs, PI detection, and low-resourced languages.

Limitations and Ethical Concerns

This is a preliminary and qualitative analysis. Our experiment featured six different prompts, three different models and three different languages, leading to $6 \times 3 \times 3$ sets of outputs. In order to fully support our claims based on the analysis of these outputs, we require evaluation and statistical analysis. This entails the manual annotation of the outputs and annotation guideline development, which was beyond the scope of this pilot study.

Another limitation of this experiment is the small number of samples, which may not reflect in style and content the types of utterances that are of interest for people wishing to use LLMs to detect personal information. Additionally, the translations into Polish were not done directly from the original, but via intermediate languages. It is also possible that more extensive tweaking of the prompt texts could lead to better performance, at least on the high-resourced language.

We also note that we aggregated the Komi-Zyrian and Komi-Permyak data without considering the differences in the models' performance between them, largely due to the fact that there are so few samples available for Komi-Permyak.

While the data that we used was sourced from openly available corpora and, therefore, likely does not pose any privacy concerns, we want to high-

light that we do not encourage the use of LLMs for PI detection without manual post-processing to ensure that no personal information is leaked, as the results are not consistent enough even for English. It is also important to keep in mind that LLMs are computationally rather heavy, and processing larger batches of text will have a noticeable carbon footprint, meaning that more lightweight solutions with similar performance may be a better choice. It is also essential to remember that LLM services hosted online may collect the users' data, so the only way to use them for PI detection without triggering privacy risks is to run them locally, which can impose high hardware requirements.

Acknowledgments

This work has been possible thanks to the funding of numerous grants from the Swedish Research Council. The first author is supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the *Centre for Linguistic Theory and Studies in Probability (CLASP)* at the University of Gothenburg. The second author's work is funded by the project *Grandma Karl is 27 years old: Automatic pseudonymization of research data* with the funding number 2022-02311 for the years 2023-2029. That author is also supported by the Swedish national research infrastructure Nationella Språkbanken, funded jointly by contract number 2017-00626 for the years 2018-2024, as well 10 participating partner institutions.

References

Flammie A Pirinen. 2024. Keeping up appearances—or how to get all Uralic languages included into bleeding edge research and software: generate, convert, and LLM your way into multilingual datasets. In *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, pages 123–131, Helsinki, Finland. Association for Computational Linguistics.

Khalid Alnajjar, Mika Hämäläinen, and Jack Rueter. 2024. Leveraging transformer-based models for predicting inflection classes of words in an endangered Sami language. In *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, pages 41–48, Helsinki, Finland. Association for Computational Linguistics.

Khalid Alnajjar, Mika Hämäläinen, Jack Rueter, and Niko Partanen. 2020. Ve'rdd. narrowing the gap between paper dictionaries, low-resource NLP and community involvement. In *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*, pages 1–6, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Rogier Blokland, Niko Partanen, and Michael Rießler. 2020. A pseudonymisation method for language documentation corpora: An experiment with spoken Komi. In *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*, pages 1–8, Wien, Austria. Association for Computational Linguistics.

Sławomir Dadas. 2019. A repository of polish NLP resources. Github.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal,

Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant

Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Mika Hämäläinen, Jack Rueter, Khalid Alnajjar, and Niko Partanen. 2023. Working towards digital documentation of Uralic languages with open-source tools and Modern NLP methods. In *Proceedings of the Big Picture Workshop*, pages 18–27, Singapore. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Beáta Megyesi, Lena Granstedt, Sofia Johansson, Julia Prentice, Dan Rosén, Carl-Johan Schenström, Gunlög Sundberg, Mats Wirén, and Elena Volodina. 2018. Learner corpus anonymization in the age of GDPR: Insights from the creation of a learner corpus of Swedish. In *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning*, pages 47–56, Stockholm, Sweden. LiU Electronic Press.

Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. 2024. The zeno's paradox of 'low-resource' languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17753–17774, Miami, Florida, USA. Association for Computational Linguistics.

Official Journal of the European Union. 2016. Consolidated text: Regulation (EU) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (general data protection regulation) (text with EEA relevance). *Official Journal*, (Document 02016R0679-20160504).

Niko Partanen, Rogier Blokland, KyungTae Lim, Thierry Poibeau, and Michael Rießler. 2018. The first Komi-Zyrian Universal Dependencies treebanks. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 126–132, Brussels, Belgium. Association for Computational Linguistics.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101.

Flammie Pirinen, Sjur Moshagen, and Katri Hiovain-Asikainen. 2023. GiellaLT — a stable infrastructure for Nordic minority languages and beyond. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 643–649, Tórshavn, Faroe Islands. University of Tartu Library.

Taido Purason, Aleksei Ivanov, Lisa Yankovskaya, and Mark Fishel. 2024a. SMUGRI-MT - machine translation system for low-resource Finno-Ugric languages. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2)*, pages 31–32, Sheffield, UK. European Association for Machine Translation (EAMT).

Taido Purason, Hele-Andra Kuulmets, and Mark Fishel. 2024b. Llms for extremely low-resource finno-ugric languages.

Jack Rueter, Niko Partanen, and Larisa Ponomareva. 2020. On the questions in developing computational infrastructure for Komi-permyak. In *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*, pages 15–25, Wien, Austria. Association for Computational Linguistics.

Piotr Rybak. 2024. Transferring BERT capabilities from high-resource to low-resource languages using vocabulary matching. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16745–16750, Torino, Italia. ELRA and ICCL.

Bahar Salehi, Dirk Hovy, Eduard Hovy, and Anders Søgaard. 2017. Huntsville, hospitals, and hockey teams: Names can reveal your location. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 116–121, Copenhagen, Denmark. Association for Computational Linguistics.

Anders Søgaard. 2022. Should we ban English NLP for a year? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5260, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nishant Subramani, Sasha Luccioni, Jesse Dodge, and Margaret Mitchell. 2023. Detecting personal information in training corpora: an analysis. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 208–220, Toronto, Canada. Association for Computational Linguistics.

Maria Irena Szawerna, Simon Dobnik, Therese Lindström Tiedemann, Ricardo Muñoz Sánchez, Xuan-Son Vu, and Elena Volodina. 2024. Pseudonymization categories across domain boundaries. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13303–13314, Torino, Italia. ELRA and ICCL.

Margaret Thomas. 2010. Names, epithets, and pseudonyms in linguistic case studies: A historical overview. *Names: A Journal of Onomastics*, 58(1):13–23.

Sunna Torge, Andrei Politov, Christoph Lehmann, Bochra Saffar, and Ziyan Tao. 2023. Named entity recognition for low-resource languages - profiting from language families. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Sixuan Wang, Junjun Muhamad Ramdani, Shuting (Alice) Sun, Priyanka Bose, and Xuesong (Andy) Gao. 2024. Naming research participants in qualitative language learning research: Numbers, pseudonyms, or real names? *Journal of Language, Identity & Education*, pages 1–14.

Jianliang Yang, Xiya Zhang, Kai Liang, and Yuenan Liu. 2023. Exploring the application of large language models in detecting and protecting personally identifiable information in archival data: A comprehensive study*. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2116–2123.

Lisa Yankovskaya, Maali Tars, Andre Tättar, and Mark Fishel. 2023. Machine translation for low-resource Finno-Ugric languages. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 762–771, Tórshavn, Faroe Islands. University of Tartu Library.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Arofat Akhundjanova, Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Erik Andersen, Matthew Andrews, Lene Antonsen, Tatsuya Aoyama, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, Bilge Nas Arıcan, Hórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Katla Ásgeirsdóttir, Deniz Baran Aslan, Cengiz Asmazoğlu, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Mariana Avelãs, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Bryan Khelven da Silva Barbosa, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Shabnam Behzad, Juan Belieni, Kepa Bengoetxea, İbrahim Benli, Yifat Ben Moshe, Ansu Berg, Gözde Berk, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Esma Fatıma Bilgin Taşdemir, Kristín Bjarnadóttir, Verena Blaschke, Rogier Blokland, Nina Böbel, Victoria Bobicev, Loïc Boizou, Johnatan Bonilla, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, António Branco, Kristina Brokaitė, Aljoscha Burchardt,

Carmen Cabeza, Natalia Cáceres Arandia, Marisa Campos, Marie Candito, Bernard Caron, Gauthier Caron, Catarina Carvalheiro, Rita Carvalho, Lauren Cassidy, Maria Clara Castro, Sérgio Castro, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Anila Çepani, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Liyanage Chamila, Claudine Chamoreau, Shweta Chauhan, Yifei Chen, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Bermet Chontaeva, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Claudia Corbetta, Daniela Corbetta, Francisco Costa, Marine Courtin, Benoît Crabbé, Mihaela Cristescu, Vladimir Cvetkoski, Netanel Dahan, Ingerid Løyning Dale, Philemon Daniel, Elizabeth Davidson, Leonel Figueiredo de Alencar, Mathieu Dehouck, Martina de Laurentiis, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Roberto Antonio Díaz Hernández, Carly Dickerson, Ariani Di Felippo, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Hoa Do, Kaja Dobrovoljc, Caroline Döhmer, Adrian Doyle, Timothy Dozat, Kira Droganova, Magali Sanches Duran, Puneet Dwivedi, Christian Ebert, Hanne Eckhoff, Masaki Eguchi, Sandra Eiche, Roald Eiselen, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Soudabeh Eslami, Farah Essaidi, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Ján Faryad, Federica Favero, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Amal Fethi, Jennifer Foster, Theodorus Fransen, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Edith Galy, Federica Gamba, Marcos Garcia, José María García-Miguel, Moa Gärdenfors, Tanja Gaustad, Efe Eren Genç, Fabrício Ferraz Gerardi, Kim Gerdes, Luke Gessler, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Gili Goldin, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Kirian Guiller, Céline Guillot-Barbance, Tunga Güngör, Vladimir Gurevich, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Takahiro Harada, Sam Hardwick, Kim Harris, Naïma Hassert, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Diana Hoefels, Petter Hohle, Nick Howell, Yidi Huang, Marivel Huerta Mendez, Jena Hwang, Takumi Ikeda, Inessa Iliadou, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Artan Islamaj, Kaoru Ito, Federica Iurescia, Sandra Jagodzińska, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Katharine Jiang, Mayank Jobanputra, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Ritván Karahóǧa, Andre Kåsen, Tolga Kayadelen, Sarveswaran Kengatharaiyer, Václava Kettnerová, Lilit Kharatyan, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Petr Kocharov, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Alexey Koshevoy, Nelda Kote, Natalia Kotsyba, Barbara Kovačić, Jolanta Kovalevskaitė, Emmanuelle Kowner, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Adrian Kuqi, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Kris Kyle, Käbi Laan, Veronika Laippala, Lorenzo Lambertino, Israel Landau, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phûông Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Lauren Levine, Cheuk Ying Li, Josie Li, Keying Li, Yixuan Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Yi-Ju Jessica Lin, Krister Lindén, Yang Janet Liu, Nikola Ljubešić, Irina Lobzhanidze, Olga Loginova, Lucelene Lopes, Edita Luftiu, Arsenii Lukashevskyi, Stefano Lusito, Anne-Marie Lutgen, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Ilya Makarchuk, Aibek Makazhanov, Francesco Mambrini, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Stella Markantonatou, Héctor Martínez Alonso, Lorena Martín Rodríguez, André Martins, Cláudia Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Maitrey Mehta, Pierre André Ménard, Gustavo Mendonça, Hilla Merhav, Tatiana Merzhevich, Paul Meurer, Niko Miekka, Emilia Milano, Aaron Miller, Yael Minerbi, Karina Mischenkova, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Lûông Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Victor Norrman, Alireza Nourian, Maria das Graças Volpe Nunes, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Hulda Óladóttir, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Noam Ordan, Petya Osenova, Robert Östling, Annika Ott, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Teresa Paccosi, Alessio Palmero Aprosio, Anastasia Panova, Thiago Alexandre Salgueiro Pardo, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Giulia Pedonese, Oggi Peeters, Angelika Peljak-Łapińska, Siyao Peng, Siyao Logan Peng, Rita Pereira, Sílvia Pereira, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria

Petrova, Andrea Peverelli, Jason Phelan, Claudel Pierre-Louis, Jussi Piitulainen, Yuval Pinter, Clara Pinto, Rodrigo Pintucci, Tommi A Pirinen, Emily Pitler, Magdalena Plamada, Barbara Plank, Alistair Plum, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Rigardt Pretorius, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Robert Pugh, Tiina Puolakainen, Christoph Purschke, Sampo Pyysalo, Peng Qi, Andreia Querido, Andriela Rääbis, Ella Rabinovich, Alexandre Rademaker, Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Joana Ramos, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Arij Riabi, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Ivan Roksandic, Norton Trevisan Roman, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Paulette Roulon, Davide Rovati, Ben Rozonoyer, Olga Rudina, Jack Rueter, Paolo Ruffolo, Kristján Rúnarsson, Rozana Rushiti, Shoval Sadde, Pegah Safari, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Konstantinos Sampanis, Stephanie Samson, Xulia Sánchez-Rodríguez, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, Marta Sartor, Albina Sarymsakova, Mitsuya Sasaki, Baiba Saulīte, Agata Savary, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Emmanuel Schang, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Sven Sellmer, Mojgan Seraji, Syeda Shahzadi, Mo Shen, Atsuko Shimada, Gyu-Ho Shin, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Maria Shvedova, Janine Siewert, Einar Freyr Sigurðsson, João Silva, Aline Silveira, Natalia Silveira, Sara Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Haukur Barri Símonarson, Kiril Simov, Dmitri Sitchinava, Ted Sither, Aaron Smith, Isabela Soares-Bastos, Per Erik Solberg, Barbara Sonnenhauser, Shafi Sourov, Rachele Sprugnoli, Vivian Stamou, Steinþór Steingrímsson, Antonio Stella, Abishek Stephen, Milan Straka, Omer Strass, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Hakyung Sung, Shingo Suzuki, Daniel Swanson, Zsolt Szántó, Chihiro Taguchi, Dima Taji, Luigi Talamo, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Mirko Tavoni, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Tarık Emre Tıraş, Sara Tonelli, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sveinbjörn Þórðarson, Vilhjálmur Þorsteinsson, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Elena Vagnoni, Sowmya Vajjala, Socrates Vak, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Uliana Vedenina, Giulia Venturi, Eric Villemonte de la Clergerie, Veronika Vincze, Anishka Vissamsetty, Natalia Vlasova, Eleni Vligouridou, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, John Wang, Jonathan North Washington, Leonie Weissweiler, Maximilan Wendt, Paul Widmer, Shira Wigderson, Sri Hartati Wijono, Vanessa Berwanger Wille, Seyi Williams, Miriam Winkler, Shuly Wintner, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Qishen Wu, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Enes Yılandiloğlu, Olcay Taner Yıldız, Zhuoran Yu, Arlisa Yuliawati, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, He Zhou, Hanzhi Zhu, Yilun Zhu, Anna Zhuravleva, Rayan Ziane, and Artūrs Znotiņš. 2024. Universal dependencies 2.15. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A Appendix: Prompt templates and examples from the parallel dataset

Here we show examples of sentences from our dataset, as well as the prompts 1 through 6, as defined in section 2 and Table 1.

(3) **Ko.**: Митяяслӧн керкаыс бокынджык грездса муköд керкаясысь , но зэв гажа местаын , неуна кыр горув лэччыштан — Эжва визувтӧ .
Pol.: Dom Mitji jest oddalony od reszty domów w wiosce , ale w bardzo miłym miejscu , dołem stromego zbocza płynie Eżwa .
Eng.: Mitya's house is remote from other houses in the village, but in a very pleasant place, slightly down a steep slope - the Ezhva flows.

(4) **Ko.**: — Эн тэрмасьӧй , Аннаыд ачыс бӧръяс , коді колӧ , — дорйис пӧдругасӧ Зоя .
Pol.: - Nie pospieszaj , Anna sama zdecyduje kto jest potrzebny - Zoja wsparła swoją przyjaciółkę .
Eng.: "Don't rush, Anna will choose who is needed herself," Zoya supported her friend.

System: You are a multilingual personal information detection tool. Personal information is information that can lead to someone in the text being reidentified, like their name, surname, middle name, patronymic, nickname, where they live, address, city, country, zip code, where they work, study, or spend a lot of their time, what unique lines or modes of transport they travel with, their age, any dates mentioned in the text, phone numbers, personal identity numbers, bank account numbers, other number sequences, e-mail addresses, urls, their work titles, education, types of family relations, information about faith, political beliefs, sexuality, ethnicity, unique achievements, etc.

User: For each token in the given text, determine whether it is a piece of personal information. Return the text with "PI" replacing every instance of personal information.

Example:
Text: I'm from Slovakia , but one of my best friends , Marie , is from Norway .
Result: I'm from PI , but one of my best friends , PI , is from PI.

Text: [PLACEHOLDER]
Result:

Figure 2: Prompt 1, [PLACEHOLDER] is replaced with an actual text.

System: You are a multilingual personal information detection tool. Personal information is information that can lead to someone in the text being reidentified, like their name, surname, middle name, patronymic, nickname, where they live, address, city, country, zip code, where they work, study, or spend a lot of their time, what unique lines or modes of transport they travel with, their age, any dates mentioned in the text, phone numbers, personal identity numbers, bank account numbers, other number sequences, e-mail addresses, urls, their work titles, education, types of family relations, information about faith, political beliefs, sexuality, ethnicity, unique achievements, etc.

User: For each token in the given text, determine whether it is a piece of personal information. Return the results in a JSON format.

Example:
Text: I'm from Slovakia , but one of my best friends , Marie , is from Norway .
Result:

```
{
   "1":{"I'm":""},
   "2":{"from":""},
   "3":{"Slovakia":"PI"},
   "4":{",":""},
   "5":{"but":""},
   "6":{"one":""},
   "7":{"of":""},
   "8":{"my":""},
   "9":{"best":""},
   "10":{"friends":""},
   "11":{",":""},
   "12":{"Marie":"PI"},
   "13":{",":""},
   "14":{"is":""},
   "15":{"from":""},
   "16":{"Norway":"PI"},
   "17":{".":""}
}
```

Text: [PLACEHOLDER]
Result:

Figure 3: One of the prompt templates used in this study. When fed to a model, [PLACEHOLDER] is replaced with an actual text.

System: You are a multilingual personal information detection tool. Personal information is information that can lead to someone in the text being reidentified, classified according to the following pattern:
<firstname_female> — women's given names
<firstname_male> — men's given names
<firstname_unknown> — given name that does not have an obvious binary gender
<surname_female> — women's surnames
<surname_male> — women's surnames
<surname_unknown> — women's surnames
<patronymic_female> — a woman's patronymic
<patronymic_male> — a man's patronymic
<street> — street names, names of squares, avenues, etc.
<city> — cities, villages, towns
<region> — regions smaller than a country
<country> — countries
<geo> — other geographical elements, such as mountains, lakes, rivers
<age> — age in digits or words

User: For each token in the given text, determine whether it is a piece of personal information. Return the text with an appropriate tag replacing every instance of personal information.

Example:
Text: I'm from Slovakia , but one of my best friends , Marie , is from Norway .
Result: I'm from <country> , but one of my best friends , <firstname_female> , is from <country>.

Text: [PLACEHOLDER]
Result:

Figure 4: Prompt 3, [PLACEHOLDER] is replaced with an actual text.

System: You are a multilingual personal information detection tool. Personal information is information that can lead to someone in the text being reidentified, classified according to the following pattern:
<firstname_female> — women's given names
<firstname_male> — men's given names
<firstname_unknown> — given name that does not have an obvious binary gender
<surname_female> — women's surnames
<surname_male> — women's surnames
<surname_unknown> — women's surnames
<patronymic_female> — a woman's patronymic
<patronymic_male> — a man's patronymic
<street> — street names, names of squares, avenues, etc.
<city> — cities, villages, towns
<region> — regions smaller than a country
<country> — countries
<geo> — other geographical elements, such as mountains, lakes, rivers
<age> — age in digits or words

User: For each token in the given text, determine whether it is a piece of personal information and assign the appropriate tag. Return the results in a JSON format.

Example:
Text: I'm from Slovakia , but one of my best friends , Marie , is from Norway .
Result:

```
{
   "1":{"I'm":""},
   "2":{"from":""},
   "3":{"Slovakia":"<country>"},
   "4":{",":""},
   "5":{"but":""},
   "6":{"one":""},
   "7":{"of":""},
   "8":{"my":""},
   "9":{"best":""},
   "10":{"friends":""},
   "11":{",":""},
   "12":{"Marie":"<firstname_female>"},
   "13":{",":""},
   "14":{"is":""},
   "15":{"from":""},
   "16":{"Norway":"<country>"},
   "17":{".":""}
}
```

Text: [PLACEHOLDER]
Result:

Figure 5: Prompt 4, [PLACEHOLDER] is replaced with an actual text.

System: You are a multilingual personal information detection tool. Personal information is information that can lead to someone in the text being reidentified, classified according to the following pattern:
<birth> — characteristics true of a person at birth, most of which are difficult or impossible to change, such as nationality, gender, caste, etc.
<society> — include characteristics that commonly develop throughout a person's life and are defined in many countries as a specially designated "status", such as immunization status.
<social> — categories corresponding to social groups such as teams or affiliations – e.g. member of the women's softball team, student of Carnegie Mellon University.
<character> — sequences of letters and numbers that can often uniquely identify a person or a small group of people; they change relatively infrequently and can therefore persist as sources of identification for years or decades – e.g. a name, surname, social security number, credit card number, IBAN, or e-mail address.
<records> — information typically consists of a persistent document or electronic analog that is not generally-available, but can allow for the (reasonable) identification of an individual – e.g. financial or health records.
<situation> — uniquely identify an individual, but that is restricted to a given context or point in time – e.g. date, time, GPS location, place of residence.

User: For each token in the given text, determine whether it is a piece of personal information. Return the text with an appropriate tag replacing every instance of personal information.

Example:
Text: I'm from Slovakia , but one of my best friends , Marie , is from Norway .
Result: I'm from <birth> , but one of my best friends , <character> , is from <birth>.

Text: [PLACEHOLDER]
Result:

Figure 6: Prompt 5, [PLACEHOLDER] is replaced with an actual text.

System: You are a multilingual personal information detection tool. Personal information is information that can lead to someone in the text being reidentified, classified according to the following pattern:
<birth> — characteristics true of a person at birth, most of which are difficult or impossible to change, such as nationality, gender, caste, etc.
<society> — include characteristics that commonly develop throughout a person's life and are defined in many countries as a specially designated "status", such as immunization status.
<social> — categories corresponding to social groups such as teams or affiliations – e.g. member of the women's softball team, student of Carnegie Mellon University.
<character> — sequences of letters and numbers that can often uniquely identify a person or a small group of people; they change relatively infrequently and can therefore persist as sources of identification for years or decades – e.g. a name, surname, social security number, credit card number, IBAN, or e-mail address.
<records> — information typically consists of a persistent document or electronic analog that is not generally-available, but can allow for the (reasonable) identification of an individual – e.g. financial or health records.
<situation> — uniquely identify an individual, but that is restricted to a given context or point in time – e.g. date, time, GPS location, place of residence.

User: For each token in the given text, determine whether it is a piece of personal information and assign the appropriate tag. Return the results in a JSON format.

Example:
Text: I'm from Slovakia , but one of my best friends , Marie , is from Norway .
Result:

```
{
   "1":{"I'm":""},
   "2":{"from":""},
   "3":{"Slovakia":"<birth>"},
   "4":{",":""},
   "5":{"but":""},
   "6":{"one":""},
   "7":{"of":""},
   "8":{"my":""},
   "9":{"best":""},
   "10":{"friends":""},
   "11":{",":""},
   "12":{"Marie":"<character>"},
   "13":{",":""},
   "14":{"is":""},
   "15":{"from":""},
   "16":{"Norway":"<birth>"},
   "17":{".":""}
}
```

Text: [PLACEHOLDER]
Result:

Figure 7: One of the prompt templates used in this study. When fed to a model, [PLACEHOLDER] is replaced with an actual text.