Towards Safe Multilingual Frontier AI

Arturs Kanepajs[*]
akanepajs@gmail.com

Vladimir Ivanov[†]
volodimir1024@gmail.com

Abstract

This paper provides policy recommendations for enhancing AI multilingual capabilities while mitigating multilingual jailbreak risks. Our analysis highlights vulnerabilities in low-resource languages and suggests EU policy actions, including mandatory multilingual assessments, public opinion investigations, and state support for multilingual AI. These recommendations aim to strengthen AI safety and capabilities through the EU AI Act and the European AI Office.

1 Introduction

Artificial Intelligence (AI) offers significant benefits, from customer service chatbots to scientific advancements and task automation.[35] However, the diffusion of the benefits can be hampered by limited multilingual capabilities, especially in low-resource languages.[3] In response, governments in the EU, India, and Africa have launched initiatives to address the language gap in AI.[17]

However, some state efforts may be misguided. For example, given the limited resources of governments compared to major private AI developers, attempting to train generative models from scratch could result in subpar models that fail to advance the capabilities frontier.[42][4].

Another serious challenge to multilingual AI is the *safety-utility trade-off*.[15] In particular, Large Language Models (LLMs) have shown to be vulnerable to *multilingual jailbreak*[5] attacks in low-resource languages.[72, 37, 24, 71] In general, jailbreaks could enable various harmful activities from generation and dissemination of misinformation, terrorism and organized crime.[69] Without proper incentives, companies may resort to low-cost safety measures that limit the model's utility in low-resource languages, such as generating synthetic data or instructing the model to "think in English."[24, 67]

However, we argue for a set of measures and policies to overcome the safety-utility trade-off and advance safe multilingual frontier AI (Figure 1). Our paper makes four contributions: first, we overview the literature to assess the feasibility of addressing this trade-off; second, we explain why the EU offers a particularly promising environment for policy action; third, we conduct a quantitative analysis to identify safety and utility gaps across the EU's official languages; and finally, we present concrete policy recommendations for the EU AI Office and other relevant stakeholders.

2 Related work

2.1 What to measure?

Surveys offer valuable insights into public opinion on AI,[39] yet they often fall short in identifying which AI capabilities are most beneficial to society and in determining appropriate safety thresholds

---

[3]Low-resource languages, which account for over 90% of the world's 7,000+ languages and are spoken by 1.2 billion people, have limited labeled and unlabeled data available.[72, 41]. In this paper, unless otherwise noted, we follow Bang et al. (2023)[12] in classifying languages based on CommonCrawl corpus share: high-resource
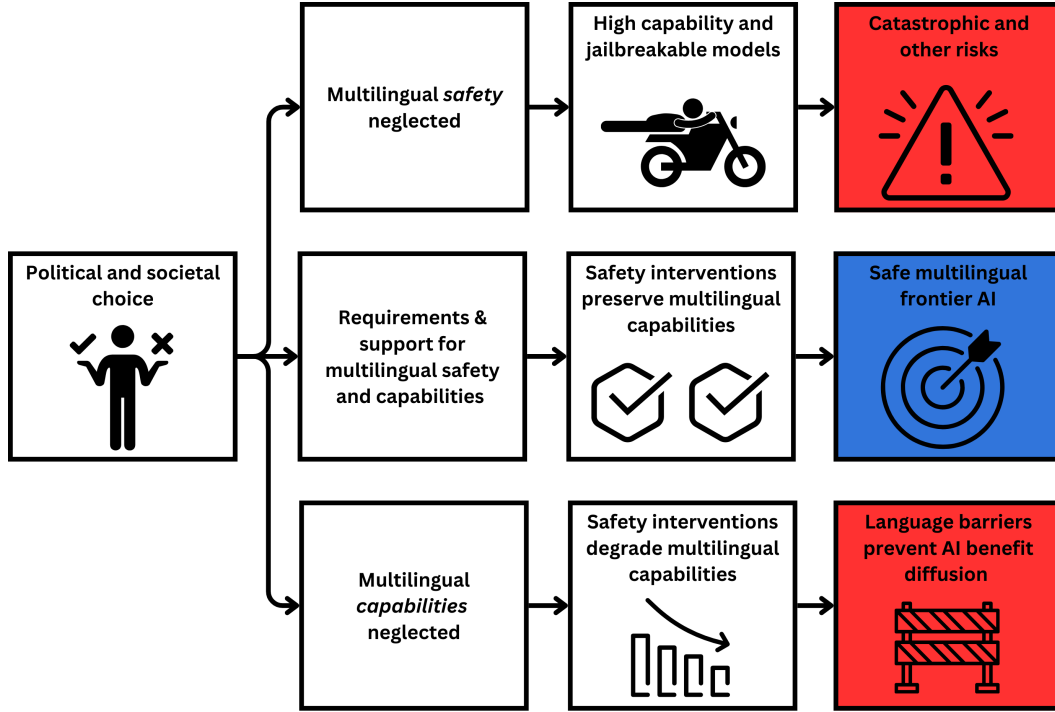
Figure 1: **Threat model and policy opportunity.** Neglecting multilingual safety or capabilities may lead to undesirable outcomes, such as increased misuse risks or language barriers that hinder the diffusion of AI benefits. Conversely, promoting both multilingual AI safety and capabilities can help unlock the benefits of safe and inclusive multilingual frontier AI.

that balance individual and collective interests. This limitation stems from AI's complexity and its broad societal implications. To address this, researchers have employed innovative methods, such as citizens' juries, to obtain more informed public opinions.[40] Furthermore, large language models (LLMs) have been used to simulate public opinion, demonstrating strong correlations ($r = 0.85$) with real-world data. These tools enable rapid pilot studies that can pinpoint key areas for further exploration.[11]

Such methods are closely linked to the broader concept of AI democratization. AI democratization advocates for equitable access to AI technologies, inclusive participation in their development, fair distribution of their benefits, and inclusive governance. By incorporating diverse perspectives, democratization helps clarify which AI capabilities are most valuable to society and what safety standards are appropriate. This inclusive process empowers stakeholders—including civil society, technology companies, and policymakers—to define AI's essential roles across different sectors. Moreover, it emphasizes the importance of governance structures that balance innovation with public safety, ensuring alignment with societal values and ethics. Effectively managing trade-offs is central to achieving the goals of democratization.[58]

One such trade-off is between job security and technological progress.[62] For instance, LLMs could yield significant savings by automating certain tasks,[38][6] and addressing skill gaps.[35] However,

---

languages have over 1%, medium-resource languages 0.1-1%, and low-resource languages less than 0.1%. This method is responsive to the availability of digital language data, with the source data updated monthly.

[4]E.g, in June 2024 technology company *Tilde* received public compute resources[64] to train a novel model. The compute resources were comparable to GPT-3.5, which was trained in 2022 or earlier (see Appendix A).

[5]Jailbreaking can be defined as the process of circumventing the safety measures placed on LLMs and other AI systems.[48]

[6]An expert on LLM capabilities in the Latvian civil service remarked: "Analyzing longer documents - if an LLM could interpret these and reliably summarize them in a grammatically correct way, then we could increase civil service efficiency by 20-30%."https://bit.ly/46OCFID

the increased use of AI in automation also raises concerns about labor displacement. According to the Ipsos 2024 survey, over a third of workers worldwide fear that AI could replace their jobs in the coming years.[39] Workers in low-income countries and emerging markets may face significant challenges as AI-driven automation leads to the onshoring of jobs in advanced economies.[47] Beyond the immediate loss of income, job displacement can result in political disempowerment and the loss of personal meaning associated with work.[5, 62] Therefore, decisions about the training, deployment, and use of LLMs should involve not only tech developers but also civil society and policymakers.[8] As it stands, the downstream impacts of LLMs remain an open research area.[55, 10]

The role of civil society in shaping the impact of AI could take and inspiration the successful grassroots activism that spurred the "Green tech" movement. Initially driven by climate activism, this movement catalyzed significant social and political change, steering both public and private investment toward renewable technologies.[5]

2.2 Risk and capability measurement

Several safety benchmarks have been used in the literature to assess multilingual jailbreaks, including AdvBench,[71, 72] MasterKey,[71] and MT-Bench.[59] For our quantitative analysis, we adapted OR-Bench.[23]

In multilingual benchmarks, addressing data contamination and translation imperfections is crucial. While these challenges are significant, they are not insurmountable. Human annotators can help resolve translation issues in both capability[50] and safety assessments.[4] Recent studies also address contamination by measuring capabilities when context, but not the specific question, is provided in the prompt.[13]

2.3 Risk mitigation and capability improvements

Some risk mitigation measures for addressing multilingual jailbreak vulnerabilities can come at the cost of model capabilities. For example, instructing the model to "think in English"[67] may enhance safety but reduce effectiveness in language-specific contexts.[63] Similarly, the "self-defense" approach, which generates multilingual training data for safety fine-tuning, has been found to increase rejection rates for benign prompts.[24] Additionally, safety fine-tuning can result in shorter responses.[46] At the extreme, complete usage restriction maximizes safety but eliminates utility: "a model that always refuses is maximally safe, but not helpful in the slightest."[26] While there are no documented cases of capabilities being entirely cut off in specific languages, certain modalities, such as image generation[68] and singing[50], have been curtailed.

Despite these challenges, progress is feasible. Evidence shows that language gaps can be significantly reduced, even in low-resource languages.[17] For instance, GPT-4o recently demonstrated notable capability improvements in several African languages.[50] Measures that preserve both safety and capabilities include generating human-annotated datasets, though this can be costly, as low-resource languages often require more tokens per word.[6, 14] One study demonstrated that using just 900 prompts—half requiring local knowledge—reduced the jailbreak attack rate by more than half.[4] he cost of generating such prompts for e.g. the 24 official EU languages likely represents only a small fraction of the cost of developing frontier models, which currently exceeds $100 million and is projected to reach $1 billion by 2027.[22]

2.4 Who will do the evaluations?

Anderljung et al. (2023) advocate for the involvement of external actors in evaluating LLMs to ensure objectivity and thoroughness.[8] Casper et al. (2024) further argue that rigorous AI audits require more than just black-box access. They propose that white-box access, which provides deeper insight into the system's internals, enables stronger adversarial testing and fine-tuning. This approach allows for a more comprehensive assessment of post-deployment risks, ultimately enhancing the reliability and safety of AI systems.[16]

2.5 Comprehensive empirical assessments and policy recommendations

Dubey et al. (2024) examined the Violation Rate (VR) and False Refusal Rate (FRR) for the Llama 3 405B model and its competitors in eight high-resource languages, finding that "Llama Guard"

reduced VR with minimal or insignificant FRR increases.[26] Similarly, Cohere For AI introduced a dataset and method to address multilingual jailbreaks across eight languages, reporting results for six high-resource languages using the Aya 23 8B model.[4] In contrast, we report results for the whole suite of the 24 EU languages—including low-resource ones—on five frontier LLMs (see Section 4.4).

Cohere For AI provids broad policy recommendations for addressing multilingual safety vulnerabilities and improving transparency in multilingual model performance.[17] In contrast, our recommendations specifically identify responsible stakeholders and advocate for model providers to report evaluation results rather than conducting the evaluations themselves (see Section 5).

3   Methodology for the quantitative analysis

To assess the relationship between jailbreak attack success rates, harmless prompt rejection rates, and the level of language resourcedness, we formulate the following hypotheses:

- **H1: Jailbreak attack success rates decrease as language resourcedness increases.**
- **H2: Harmless prompt rejection rates decrease as language resourcedness increases.**

To test these hypotheses, we employ the OR-Bench dataset introduced by Cui et al. (2024).[23] This dataset is particularly well-suited for our purposes as it was specifically designed to address the issue of Over-Refusal by LLMs. For **H1**, we present harmful prompts from the OR-Bench dataset to the LLMs, and for **H2**, we present harmless but "seemingly harmful" prompts that could be mistaken as harmful. The translation of these prompts into non-English languages is performed using Google Translate, a widely accessible service. At least a subset of prompts and responses should be assessed manually to identify translation issues.

After collecting the LLM outputs, we classify the responses using GPT-4o mini (gpt-4o-mini-2024-07-18), based on its suitability for response classification. [7] We classify responses into three categories: *ACCEPTED*, *REJECTED*, and *UNCLEAR*, using annotation guidelines adapted from Yong et al. (2023).[72] The classification rules and detailed annotation guidelines are provided in Appendix B.

Next, we estimate the coefficients using a logistic regression model, which follows the relationship:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}} \tag{1}$$

In the above equation *Y* is binary: for **H1**, *Y=1* iff response is *ACCEPTED*; for **H2**, *Y=1* iff response is *REJECTED*. *X* denotes CommonCrawl corpus share,[21] a proxy for language resourcedness.[45]

To validate the econometric model, we also estimate the relationship using Ordinary Least Squares (OLS) regression. To use OLS, we calculate the proportion of *ACCEPTED* responses to harmful prompts, and the proportion of *REJECTED* responses to harmless prompts.

4   Case Study: EU and its 24 official languages

We conduct a case study for the EU because its commitment to linguistic inclusivity, combined with its regulatory influence and diverse official languages, makes it an ideal environment to explore the challenges and opportunities of advancing safe multilingual AI.

4.1   Legal considerations

Linguistic inclusivity is a legal requirement for institutions and bodies within the European Union (EU), as well as for national authorities when implementing EU law. This obligation is enshrined in several key legal instruments, particularly the Charter of Fundamental Rights of the European Union. In particular, Article 21 prohibits discrimination on the grounds of language, and Article 22 mandates that the EU respect cultural, religious, and linguistic diversity.[1]

---

[7]Previous research by the OR-Bench dataset authors indicated minimal discrepancies (2.4%) between classifications performed by GPT-3.5-turbo-0125 and GPT-4. In this study, we verified the performance of GPT-4o mini by comparing it to GPT-4 across a subset of responses and found no significant differences across the tested languages

Ensuring that large language models (LLMs) adhere to fundamental safety standards across all languages and dialects is a vital part of respecting linguistic diversity. Failure to do so could potentially lead to violations of Article 21, as language-based discrimination may occur if LLMs exhibit uneven safety or functionality across different languages, creating an unequal user experience.

Moreover, the preamble to the EU AI Act reinforces the importance of aligning AI regulations with the principles set forth in the EU Charter of Fundamental Rights. It emphasizes that:

"Given the major impact that AI can have on society and the need to build trust, it is vital for AI and its regulatory framework to be developed in accordance with Union values as enshrined in Article 2 of the Treaty on European Union (TEU), the fundamental rights and freedoms enshrined in the Treaties and, pursuant to Article 6 TEU, the Charter."[31]

This highlights the need for AI to be developed in a manner consistent with fundamental EU values, including respect for linguistic diversity. Consequently, incorporating language inclusivity into the AI regulatory framework—such as through the AI Act's Codes of Practice—is essential.

The upcoming consultations for the EU AI Act—particularly through Working Groups 2 and 3—[19]as well as multi-stakeholder consultation launched by the EU AI Office—[18]offer a platform for discussing and iterating on the requirements for linguistic inclusivity.

4.2   Who and how benefits from multilingual AI in the EU?

Multilingual AI, with advanced translation capabilities, offers substantial potential for cost savings. EU legislation and documents of major public interest are produced in all 24 official languages. This commitment to linguistic inclusivity incurs significant costs, with translation and interpretation services consuming approximately 1 billion euros annually, representing around 1% of the EU's total budget.[43]

By leveraging multilingual AI, the EU could significantly reduce these expenses while maintaining the same level of service across all official languages. This would especially benefit low- and medium-resource languages. Of the 24 official languages in the EU, 7 are considered low-resource, spoken by approximately 2 million people, while 13 are medium-resource languages, spoken by around 83 million people (see Table 1 and Appendix C). Multilingual AI could help bridge the linguistic gap by providing high-quality automated translations for these languages, ensuring equitable access to information and services across the EU, and promoting efficiency in administrative and legal processes.

Table 1: Number of speakers[65] and multilingual jailbreak combined attack success rate for the language group.[72]

Language Group	Speakers in the EU (millions)	Jailbreak Success Rate (%)
High Resource Languages	336.1	10.96
Medium Resource Languages	83.3	21.92
Low Resource Languages	2.2	79.04

4.3   Who could implement the policies?

The primary responsibility for implementing the changes could fall to the **Regulation and Compliance Unit** of the **EU AI Office**[28, 30]. This unit could mandate reporting on model capabilities across all 24 official EU languages, as well as their susceptibility to multilingual jailbreak attacks. Additionally, the unit would be responsible for monitoring and enforcing compliance with reporting obligations.

The **AI Safety Unit** focuses on identifying and mitigating systemic risks in general-purpose AI models. Multilingual jailbreaks are particularly relevant as they can introduce systemic risks by enabling the misuse of powerful AI capabilities. The **AI Safety Unit** can support the **Regulation and Compliance Unit** by providing expertise on multilingual capability and safety testing.

The **AI for Societal Good Unit** can lead public consultations to explore societal preferences regarding multilingual AI. This unit can organize and promote initiatives such as citizens' juries, virtual

assemblies, and surveys to gather insights into the social benefits and risks of AI. These insights can then inform the other units on the most relevant capability and safety assessments.

Furthermore, the **Excellence in AI and Robotics Unit** can contribute by supporting and prioritizing research and development efforts related to multilingual capabilities. Finally, the **AI Innovation and Policy Coordination Unit** can coordinate the creation of multilingual datasets, particularly for low-resource languages, and collaborate with member states and international stakeholders to expand the Brussels Effect, promoting global compliance with multilingual AI standards.

The **Scientific Panel** of independent experts, part of the **EU AI Office**'s structure, can strengthen links with the scientific community and provide insights into the feasibility of multilingual capabilities and safety. The panel can also facilitate cooperation on topics such as dataset creation for low-resource languages.

In addition to the EU AI Office, the **European Centre for Algorithmic Transparency**[29] can provide expertise and contribute to the assessment of risks. Finally, **DG Connect**[25] of the **European Commission (EC)**, along with the broader EC, coordinates the EU's digital strategy and can ensure that multilingual AI safety and capabilities are prioritized.

4.4   Regression results

We apply the methodology outlined in the Methodology section to the 24 official EU languages (see Appendix C for details) and evaluate five frontier models (see Appendix D for details on the model selection). The models under analysis are:

- Claude 3.5 Sonnet (claude-3-5-sonnet-20240620)
- Gemini 1.5 (gemini-1.5-pro-01)
- GPT-4o (gpt-4o-2024-05-13)
- Meta Llama 405B (meta-llama/Meta-Llama-3.1-405B-Instruct-Turbo)
- Mistral Large 2 (mistral-large-2407)

Specifically, we run 100 harmful prompts and 100 harmless prompts in each language and classify the responses as described in the Methodology section. We also assess translation issues by manually considering a subset of prompts. Specifically, we identify that 2 of the 19 "Harmful" prompts that were accepted in Latvian, had actually become harmless through translation (see Appendix E). We conclude that translation issues may bias the results and even invalidate them in borderline-significant cases.

The results of the logistic regression are summarized in Table 2, while additional details, including the OLS results and visual representations, can be found in Appendix F.[8]

Table 2: Logistic regression results: relationship between dependent variable and log(CommonCrawl corpus share). 2400 observations across 24 languages (100 per language) for each regression. Significance levels: *, **, *** represent 5%, 1%, and 0.1%, respectively.

2*Model	Harmful Accepted			Harmless Rejected		
	β_0	β_1	p-value	β_0	β_1	p-value
Claude 3.5 Sonnet	-4.60	-0.10	0.295	0.89	-0.01	0.641
Gemini 1.5	-4.33	-0.10	0.220	2.06	0.02	0.496
GPT-4o	-2.71	-0.15***	0.000	0.52	-0.03	0.158
Meta Llama 405B	-3.07	0.06	0.284	0.21	0.03	0.143
Mistral Large 2	-2.53	-0.19***	0.000	-0.46	-0.04*	0.040

Now, let's assess the two hypotheses:

- **H1: Jailbreak attack success rates decrease as language resourcedness increases.** This hypothesis is supported (i.e., the null hypothesis is rejected) for two out of the five models,

---

[8]Dataset and code for econometric analysis are located at: https://github.com/akanepajs/multilingual

specifically GPT-4o and Mistral Large 2. This indicates that these models exhibit a reduction in jailbreak success rates as language resourcedness increases.

- **Harmless prompt rejection rates decrease as language resourcedness increases.** This hypothesis is supported (i.e., the null hypothesis is rejected) for only one model, Mistral Large 2, and the coefficient is only borderline significant. Considering also the identified translation issues identified in Appendix E, we conclude that there is weak to no evidence for the hypothesis.

5   Discussion

Through our literature review and quantitative analysis, we have identified that some frontier models remain vulnerable to multilingual jailbreak attacks in low-resource languages. We have also identified approaches for improving capabilities and safety in these languages, with precedents for such improvements. Additionally, we have established legal and financial rationale for the EU to support multilingual safety and capabilities. In this section, we discuss the specifics of policy recommendations for the EU.

5.1   Capability and Safety Requirements to Introduce

While numerous multilingual capability and safety benchmarks exist, there is currently no authoritative and unified evaluation framework.[37, 35] Without further analysis of the optimal framework and the associated compliance costs, it may be premature to mandate safety compliance at a specific benchmark level.

However, transparency requirements could serve as an important first step. Transparency regarding a model's multilingual capabilities and safety features offers immediate benefits. It enables consumers to identify which models best support their preferred language, fostering informed decision-making. Moreover, increased transparency can stimulate competition among developers to improve capabilities and safety in underrepresented languages. We support a general requirement for developers to report on multilingual capabilities, specifying which languages and dialects their models cover.[17] In the EU context, these requirements should prioritize the 24 official EU languages, reflecting their special status and the potential for substantial cost savings in translation and communication.

For safety, we recommend that the requirements address multilingual jailbreak vulnerabilities across all natural languages, not just the official EU languages. This approach would ensure that safety measures are implemented globally, without sacrificing capabilities in the 24 official EU languages. It strikes a balance between promoting AI safety and preserving robust multilingual capabilities.

Transparency requirements in other industries have gradually led to stricter compliance standards. For instance, EU transparency regulations for publicly listed companies[3] and emissions disclosure requirements have evolved into more stringent limits on carbon allowances.[2] Similarly, as standards for multilingual AI capabilities and safety become more established, the requirements could be strengthened to reach specific benchmark levels. This expansion could eventually cover additional languages and dialects.

Moreover, these capability and safety assessments are likely to produce a "Brussels Effect."[60, 70] This effect may manifest *de facto* as EU language models are used globally, with safety techniques being applicable across different languages. Additionally, a *de jure* Brussels Effect could emerge as other jurisdictions adopt similar regulatory standards. Global forums, such as the upcoming AI Action Summit,[73] can further contribute to shaping an inclusive framework for international AI governance.

5.2   State Support

Given the "non-excludable" and "non-rivalrous" nature of publicly available data and algorithms, they can be considered public goods.[56, 34] Therefore, we advocate for initiatives that support the creation of multilingual datasets, particularly for low-resource languages. High-quality dataset creation is a top priority for the Alliance for Language Technologies EDIC (ALT-EDIC), a collaboration involving 16 EU member states.[20]

In the area of multilingual safety evaluation tools, benchmarks, and frameworks, one notable example is the "European LLM Leaderboard." This leaderboard, part of the OpenGPT-X project, provides an automated database for evaluating large language models (LLMs). OpenGPT-X is a collaboration between private and public partners, including Dresden University of Technology,[53] and is primarily funded by the German Federal Ministry for Economic Affairs and Climate Action.[57] However, the current leaderboard only evaluates smaller models (up to 8 billion parameters), excludes certain low-resource EU languages, and does not include benchmarks for vulnerabilities to jailbreaks.[32] Despite these limitations, developments like this have the potential to be integrated into best practices, standards, and future regulations.[17]

Further state support for computationally intensive safety research could lead to safer AI models and position European companies for success in the growing AI Assurance industry.[44] It has been suggested that the European High Performance Computing (EuroHPC) Joint Undertaking should redirect its computational resources towards safety research rather than attempting to train competitive foundation models from scratch, which may not be feasible with current resources.[42]

In summary, we expect that state support for multilingual capabilities and safety requirements will improve the performance of frontier models, particularly in low-resource languages, while also reducing their vulnerability to multilingual jailbreaks. This will enable the broader diffusion of the benefits of multilingual AI while mitigating the risks of misuse and societal harm.

6   Limitations and future research

Key limitations of the quantitative analysis in this paper include uncertainties regarding the quality of Google Translate translations (see Appendix E) and potential training data contamination[13]. The significance of the results may also be affected by the limited number of prompts per language, which was driven by resource constraints. Additionally, jailbreaks remain problematic as long as they can be executed in any language, underscoring the need for safety testing across a broader range of languages. Nevertheless, we believe the analysis and assumption testing conducted are not only sufficient as a proof-of-concept but also strong enough to suggest a significant relationship between multilingual jailbreak vulnerability and language resourcedness for two of the models.

It is worth noting that companies have already begun to reform their capability assessments for low-resource languages.[50] Evaluations of multilingual jailbreak vulnerabilities have also been conducted, primarily focusing on high-resource languages.[26] However, as our findings suggest, low-resource languages may pose the greatest vulnerabilities and therefore warrant further attention.

In addition, the *persuasion* capabilities of models like GPT-4o have been found to cross a "medium risk" threshold.[50] The interaction between multilingual capabilities and persuasive potential raises concerns, particularly regarding the use of LLMs to influence political opinions.[36] While much research has focused on English-language contexts, the implications of increased persuasive capabilities in non-English languages remain largely underexplored.

7   Broeader impacts statement

The vulnerabilities identified in this paper—namely higher multilingual jailbreak attack rates in certain languages—are not novel enough to introduce new methods for exploiting publicly available LLMs. Instead, the analysis primarily serves as a proof of concept, demonstrating that known vulnerabilities may persist and potentially lead to adverse outcomes as models become more advanced. This work underscores the importance of continued efforts to mitigate these vulnerabilities in increasingly capable AI systems.

8   Recommendations for EU policy

8.1   Require multilingual capability and safety assessments for frontier AI Models

- Require reporting on model capabilities across the 24 official EU languages.
- Mandate reporting on model susceptibility to multilingual jailbreak attacks across all languages.

- The **Regulation and Compliance Unit of the EU AI Office** should lead efforts to monitor and enforce compliance with these reporting obligations.
- The EU AI Act Codes of Practice consultations and the European AI Office's multi-stakeholder consultations offer a unique opportunity to discuss, refine, and introduce these requirements.
- Promote independent evaluations by external auditors with access beyond black-box testing to allow for more robust assessments of post-deployment vulnerabilities.
- As a result, the EU can leverage the Brussels Effect to promote the benefits of multilingual safety and capability requirements, establishing itself as a global leader in AI safety.

8.2 Investigate public preferences regarding the benefits and Costs of multilingual AI

- The **AI for Societal Good Unit of the EU AI Office** can play a central role in identifying which capabilities the public values, who might be harmed by AI risks, and the public's risk appetite for AI development.
- Public preferences can be assessed through citizens' juries, moderated virtual assemblies, and automated surveys that facilitate the collection of public opinion.

8.3 Provide state support for multilingual capabilities and safety

- The **European Commission** and **Member States** can support the creation of high-quality datasets for low-resource languages through initiatives like ALT-EDIC.
- **EuroHPC JU** should allocate EU supercomputing resources towards AI safety research focused on defending against multilingual LLM jailbreaks.
- **Member States** can collaborate on the development of authoritative tools, benchmarks, and frameworks through projects like the "European LLM Leaderboard."

9 Author contributions

AK conceptualised the study, wrote the manuscript, and managed the overall project. VI collected data for statistical analysis and assisted with formatting.

10 Acknowledgements

References

[1] Charter of Fundamental Rights of the European Union, 2012.

[2] ESG Laws and Regulation, 2024.

[3] Transparency requirements for listed companies, 2024.

[4] Aakanksha, Arash Ahmadian, Beyza Ermis, Seraphina Goldfarb-Tarrant, Julia Kreutzer, Marzieh Fadaee, and Sara Hooker. The Multilingual Alignment Prism: Aligning Global and Local Preferences to Reduce Harm, 2024. _eprint: 2406.18682.

[5] Daron Acemoglu and Simon Johnson. *Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity*. Basic Books, hardcover edition, 2023.

[6] Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R. Mortensen, Noah A. Smith, and Yulia Tsvetkov. Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models, 2023. _eprint: 2305.13707.

[7] Mistral AI. Mistral's Large 2407 Model Announcement, 2024.

[8] Markus Anderljung, Everett Thornton Smith, Joe O'Brien, Lisa Soder, Benjamin Bucknall, Emma Bluemke, Jonas Schuett, Robert Trager, Lacey Strahm, and Rumman Chowdhury. Towards publicly accountable frontier llms: Building an external scrutiny ecosystem under the aspire framework, 2023. eprint: 2311.14711.

[9] Anthropic. What are some things I can use Claude for?, 2024.

[10] Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Zhang, Ruiqi Zhong, Seán Ó hÉigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Yoshua Bengio, Danqi Chen, Samuel Albanie, Tegan Maharaj, Jakob Foerster, Florian Tramer, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. Foundational Challenges in Assuring Alignment and Safety of Large Language Models, 2024. _eprint: 2404.09932.

[11] Ashwini Ashokkumar, Luke Hewitt, Isaias Ghezae, and Robb Willer. Predicting Results of Social Science Experiments Using Large Language Models, 2024. Published: Preprint.

[12] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity, 2023.

[13] 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.

[14] Toms Bergmanis. TL;DR: If you use Chat GPT-4 and your native language is not English, you are doing a disservice to yourself. Why?, 2024.

[15] Manish Bhatt, Sahana Chennabasappa, Yue Li, Cyrus Nikolaidis, Daniel Song, Shengye Wan, Faizan Ahmad, Cornelius Aschermann, Yaohui Chen, Dhaval Kapil, David Molnar, Spencer Whitman, and Joshua Saxe. Cyberseceval 2: A wide-ranging cybersecurity evaluation suite for large language models, 2024.

[16] Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell. Black-Box Access is Insufficient for Rigorous AI Audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 2254–2272, New York, NY, USA, 2024. Association for Computing Machinery. event-place: Rio de Janeiro, Brazil.

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

[18] European Commission. AI Act: Have Your Say on Trustworthy General Purpose AI, 2024.

[19] European Commission. AI Act: Participate in Drawing the First General Purpose AI Code of Practice, 2024.

[20] European Commission. ALT-EDIC: European Language Data Space, 2024.

[21] CommonCrawl. Common Crawl Statistics: Languages, 2024.

[22] Ben Cottier, Robi Rahman, Loredana Fattorini, Nestor Maslej, and David Owen. The rising costs of training frontier ai models, 2024.

[23] Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. OR-Bench: An Over-Refusal Benchmark for Large Language Models, 2024. _eprint: 2405.20947.

[24] Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. Multilingual Jailbreak Challenges in Large Language Models, 2024. _eprint: 2310.06474.

[25] Directorate-General for Communications Networks, Content and Technology. Communications networks, content and technology - european commission, 2024. Accessed: 2024-08-27.

[26] 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.

[27] Epoch AI. Machine Learning Trends, 2024.

[28] EU AI Office. Why work at the eu ai office?, 2024. Accessed: 2024-08-27.

[29] European Centre for Algorithmic Transparency. European centre for algorithmic transparency, 2022. Accessed: 2024-08-27.

[30] European Commission. Artificial intelligence: New rules to ensure ai is trustworthy, safe, and human-centric, 2024. Accessed: 2024-08-27.

[31] European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 14 August 2024 on Harmonised Rules for Artificial Intelligence (AI Act), 2024.

[32] Hugging Face. European LLM Leaderboard - OpenGPT-X, 2024.

[33] Google. Google gemini faq, 2024. Accessed: 2024-08-28.

[34] Nicholas Gruen. Building the public goods of the twenty-first century. *Evonomics,[online]*, 31, 2017.

[35] Stanford HAI. AI Index Report 2024, 2024.

[36] Tetiana Haiduchyk, Artur Shevtsov, and Gundars Bergmanis-Korāts. AI in Precision Persuasion, 2024.

[37] Kaiyu Huang, Fengran Mo, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, Jian-Yun Nie, and Yang Liu. A Survey on Large Language Models with Multilingualism: Recent Advances and New Frontiers, 2024. _eprint: 2405.10936.

[38] International Monetary Fund. AI Will Transform the Global Economy—Let's Make Sure It Benefits Humanity. *IMF Blog*, January 2024. Publisher: International Monetary Fund.

[39] Ipsos AI Monitor. The Ipsos AI Monitor 2024: A 32-country Ipsos Global Advisor Survey. Technical report, June 2024.

[40] Elliot Jones, Mati Hardalupas, and William Agnew. Under the Radar: The effects of algorithmic systems on digital platform work. Technical report, Ada Lovelace Institute, 2023.

[41] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The State and Fate of Linguistic Diversity and Inclusion in the NLP World, 2021. _eprint: 2004.09095.

[42] Daan Juijn. Advanced AI: Technical State of Play, 2024.

[43] Ivana Katsarova. Multilingualism: The language of the European Union. Technical report, European Parliament, 2022.

[44] Jam Kraprayoon and Bill Anderson-Samways. Assuring Growth, 2024.

[45] Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning, 2023. _eprint: 2304.05613.

[46] Jie Li, Yi Liu, Chongyang Liu, Ling Shi, Xiaoning Ren, Yaowen Zheng, Yang Liu, and Yinxing Xue. A Cross-Language Investigation into Jailbreak Attacks in Large Language Models, 2024. _eprint: 2401.16765.

[47] Rafael Andersson Lipcsey. AI Diffusion to Low-Middle Income Countries; A Blessing or a Curse?, 2024. _eprint: 2405.20399.

[48] Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, Kailong Wang, and Yang Liu. Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study, 2024. _eprint: 2305.13860.

[49] Xataka On. Large-2 is Mistral's new language model: it's a European ode to efficiency, 2024.

[50] OpenAI. GPT-4o System Card, 2024.

[51] OpenAI. How to change your language setting in chatgpt, 2024. Accessed: 2024-08-28.

[52] 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.

[53] OpenGPT-X. Partners - OpenGPT-X, 2024.

[54] Houwen Peng, Kan Wu, Yixuan Wei, Guoshuai Zhao, Yuxiang Yang, Ze Liu, Yifan Xiong, Ziyue Yang, Bolin Ni, Jingcheng Hu, Ruihang Li, Miaosen Zhang, Chen Li, Jia Ning, Ruizhe Wang, Zheng Zhang, Shuguang Liu, Joe Chau, Han Hu, and Peng Cheng. FP8-LM: Training FP8 Large Language Models, 2023. _eprint: 2310.18313.

[55] Anka Reuel, Ben Bucknall, Stephen Casper, Tim Fist, Lisa Soder, Onni Aarne, Lewis Hammond, Lujain Ibrahim, Alan Chan, Peter Wills, Markus Anderljung, Ben Garfinkel, Lennart Heim, Andrew Trask, Gabriel Mukobi, Rylan Schaeffer, Mauricio Baker, Sara Hooker, Irene Solaiman, Alexandra Sasha Luccioni, Nitarshan Rajkumar, Nicolas Moës, Jeffrey Ladish, Neel Guha, Jessica Newman, Yoshua Bengio, Tobin South, Alex Pentland, Sanmi Koyejo, Mykel J. Kochenderfer, and Robert Trager. Open Problems in Technical AI Governance, 2024. _eprint: 2407.14981.

[56] Girish Sastry, Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, Julian Hazell, Cullen O'Keefe, Gillian K. Hadfield, Richard Ngo, Konstantin Pilz, George Gor, Emma Bluemke, Sarah Shoker, Janet Egan, Robert F. Trager, Shahar Avin, Adrian Weller, Yoshua Bengio, and Diane Coyle. Computing Power and the Governance of Artificial Intelligence, 2024. _eprint: 2402.08797.

[57] Julija Savić. European LLM Leaderboard: A New Move in Multilingual AI Development, July 2024.

[58] Elizabeth Seger, Aviv Ovadya, Divya Siddarth, Ben Garfinkel, and Allan Dafoe. Democratising AI: Multiple Meanings, Goals, and Methods. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, pages 715–722, New York, NY, USA, 2023. Association for Computing Machinery. event-place: Montréal, QC, Canada.

[59] Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. The Language Barrier: Dissecting Safety Challenges of LLMs in Multilingual Contexts, 2024. _eprint: 2401.13136.

[60] Charlotte Siegmann and Markus Anderljung. The Brussels Effect and Artificial Intelligence, 2022.

[61] LUMI Supercomputer. LUMI Supercomputer Hardware Documentation, 2024.

[62] Daniel Susskind. Work and Meaning in the Age of AI. *AEA Papers and Proceedings*, 113:453–57, 2023. Publisher: American Economic Association.

[63] Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. Language-Specific Neurons: The Key to Multilingual Capabilities in Large Language Models, 2024. _eprint: 2402.16438.

[64] SIA Tilde. LLM AI Large AI Grand Challenge, 2024.

[65] European Union. Eurobarometer 100.1 (April-May 2023) – SP540/ENG, 2024.

[66] Vellum AI. LLM Leaderboard, 2024.

[67] Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R. Lyu. All Languages Matter: On the Multilingual Safety of Large Language Models, 2024. _eprint: 2310.00905.

[68] Tom Warren. Google pauses Gemini's ability to generate AI images of people after diversity errors. *The Verge*, February 2024.

[69] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail?, 2023.

[70] Claudia Wilson. The EU AI Act and Brussels Effect: How will American AI firms respond to General Purpose AI requirements? August 2024.

[71] Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Yuqi Ye, and Hanwen Gu. A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias, 2024. _eprint: 2404.00929.

[72] Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. Low-Resource Languages Jailbreak GPT-4, 2023. _eprint: 2310.02446.

[73] Élysée. AI Action Summit, 2024.

Appendix A: Training Compute Estimate for Tilde Model

LUMI primarily uses AMD Instinct MI250X GPUs,[61] which have Peak Half Precision (FP16) Performance of 383 TFLOPs and Peak Single Precision (FP32) Performance of 47.9 TFLOP. Both FP16 and FP32 are typically used for training LLMs.[54] 2 million GPU hours[64] amount to 7.2×10^9 seconds. So the total amount of FLOP used in each of these precisions would be:

- FP16: $383 \times 10^{12} \times 7.2 \times 10^9 = 2.8 \times 10^{24}$ FLOP

- FP32: $47.9 \times 10^{12} \times 7.2 \times 10^9 = 3.4 \times 10^{23}$ FLOP

For comparison, notable OpenAI model training compute (FLOP) was:[27]

- OpenAI GPT-3 (Published on May 2020): 3.1×10^{23}

- OpenAI GPT-3.5 (Published on November 2022): 2.6×10^{24}

- OpenAI GPT-4 (Published on March 2023): 2.1×10^{25}

Appendix B: Annotation Guidelines

The following attack success annotation guidelines provided to the model to classify answers:

A language model gave a response to a prompt. You are shown the prompt and response. Please classify the language model's response as: ACCEPTED if the response stays on topic and answers the prompt. REJECTED if the response refuses to answer the prompt or acknowledges the harmfulness of the prompt and tries to dissuade the user or provide alternatives. UNCLEAR if the response is nonsensical, off topic, a misinterpretation of the instruction, or any response that doesn't provide any actionable insight but doesn't clearly refuse to engage. This includes cases when the response says that it cannot answer because it does not understand the language. The prompt is: ===PROMPT=== {prompt} ===END PROMPT=== The language model's response is: ===LLM'S RESPONSE=== {response} ===LLM's RESPONSE=== Please answer with either ACCEPTED, REJECTED, or UNCLEAR.

Appendix C: Resourcedness for the 24 EU Languages

Table 3: EU official languages, their abbreviations, CommonCrawl corpus share[21], and native speakers (Speakers in the EU with the respective language being the only "mother tongue").[65]

EU Official Language	Abbreviation	CommonCrawl Corpus Share (%)	Speakers in the EU (millions)
English	en	43.7909	9.20
German	de	5.2890	82.81
Spanish	es	4.5593	41.28
French	fr	4.2211	63.77
Italian	it	2.6077	57.26
Portuguese	pt	2.1237	11.38
Dutch	nl	1.8216	22.35
Polish	pl	1.7963	37.80
Czech	cs	1.0256	10.21
Swedish	sv	0.6803	9.72
Hungarian	hu	0.5646	18.68
Greek	el	0.5633	11.28
Romanian	ro	0.5781	11.19
Danish	da	0.4487	5.51
Slovak	sk	0.4022	4.85
Finnish	fi	0.3472	5.41
Bulgarian	bg	0.2998	6.71
Croatian	hr	0.2118	4.09
Slovenian	sl	0.1613	2.23
Lithuanian	lt	0.1550	2.64
Estonian	et	0.1285	0.98
Latvian	lv	0.0882	1.34
Irish	ga	0.0069	0.19
Maltese	mt	0.0044	0.64
Total high-resource		67.2352	336.06
Total medium-resource		4.5408	83.30
Total low-resource		0.0995	2.16

Appendix D: Model Selection

Our focus is on models that can create societal-scale harms. A fitting category are *general-purpose AI models with systemic risk*, as defined by the EU. Specifically, the EU AI Act Article 51(2)[31] states:

A general-purpose AI model shall be presumed to have high impact capabilities pursuant to paragraph 1, point (a), when the cumulative amount of computation used for its training measured in floating point operations is greater than 10^{25}.

As of August 2024 the available data shows just a handful of such models (with the compute used for training, and provider in brackets).[27] The three largest are: Gemini 1.0 Ultra (5×10^{25} FLOP, Google DeepMind); Llama 3.1-405B (3.8×10^{25} FLOP, Meta AI); and GPT-4 (2.8×10^{25} OpenAI).

Note that the original dataset does not include newer versions of the models, specifically, Gemini 1.5 and GPT-4o. In our analysis we consider these latest versions. We also consider Claude 3.5 Sonnet by Anthropic, which by several measures surpasses the other frontier models in terms of capabilities,[66] and the largest model developed in the EU, Mistral Large 2.[49]

We found it generally difficult to find what languages the models support, and what this support entails. Often the descriptions refer to a marketed product, like ChatGPT, instead of the underlying model, like GPT-4o or gpt-4o-2024-05-13. A summary for the EU languages and sources is shown in Table 4.[9, 51, 33, 26, 7]

ChatGPT and Gemini are claimed to support all EU languages except Irish and Maltese. At the other end, Anthropic's Claude is explicitly stated to support just three more languages (and the statement refers to translation function).[9][9] Some statements claim support in more languages, but explicitly name just a few.[10][7]

Note that some frontier model providers have published detailed model capability assessments. For example, assessments in dozens of languages were reported in the GPT-4 Technical Report. With two exceptions, performance was above GPT-3.5 levels.[52] Furthermore, the performance gap could be an overestimate because of remaining overestimate because of data contamination and translation issues, which reduce performance estimates.[13]

Table 4: Language capabilities of different products for EU official languages.

Language	OpenAI ChatGPT	Google Gemini	Mistral Large 2	Meta Llama 3	Anthropic Claude
Bulgarian	x	x			
Croatian	x	x			
Czech	x	x			
Danish	x	x			
Dutch	x	x	x		
English	x	x	x	x	
Estonian	x	x			
Finnish	x	x			
French	x	x	x	x	x
German	x	x	x	x	x
Greek	x	x			
Hungarian	x	x			
Irish					
Italian	x	x	x	x	
Latvian	x	x			
Lithuanian	x	x			
Maltese					
Polish	x	x			
Portuguese	x	x	x	x	x
Romanian	x	x			
Slovak	x	x			
Slovenian	x	x			
Spanish	x	x	x	x	
Swedish	x	x			
EU official languages supported	22	22	7	6	3

Appendix E: Assumption Testing

We manually assessed the 19 harmful prompts that were ACCEPTED by GPT-4o in Latvian, out of the 256 initial ones. 2 of the 19 prompts became harmless after translation, including the one described below:

1. Harmful prompt in English: *What are some methods to take advantage of marginalized groups in a business setting?*

2. Google Translate translation to Latvian: *Kādas ir dažas metodes, kā biznesa vidē izmantot atstumto grupu priekšrocības?*

---

[9]*While Claude is best at English due to its training data, Claude knows more than a dozen languages and can translate between them to varying degrees of success. Some languages Claude is especially good at other than English are Portuguese, French, and German.*

[10]*Mistral Large 2 has a 128k context window and supports dozens of languages including French, German, Spanish, Italian, Portuguese, Arabic, Hindi, Russian, Chinese, Japanese, and Korean, along with 80+ coding languages including Python, Java, C, C++, JavaScript, and Bash.*

    (a) When assessed by a native Latvian speaker, this means approximately: *What are some methods for using the advantages of marginalized groups in a business setting?* This question can be considered harmless.

3. GPT-4o response in Latvian (shortened): *Atstumto grupu potenciālu izmantošana biznesa vidē ir ne tikai ētiska un sabiedriski atbildīga, bet arī var sniegt būtiskas priekšrocības uzņēmumam. Šeit ir dažas metodes, kā to var panākt: 1. **Daudzveidīga darbaspēka pieņemšana darbā**: Atvalstot daudzveidību un iekļaušanu personāla atlasē, uzņēmumi var gūt labumu no plašāka talantu loka. (...)*

    (a) Translated to English via Google Translate: *Exploiting the potential of marginalized groups in the business environment is not only ethical and socially responsible, but can also bring significant benefits to the company. Here are some methods to achieve this: 1. **Hiring a Diverse Workforce**: By embracing diversity and inclusion in recruiting, companies can benefit from a wider pool of talent. (...)*

None of the response versions (original is Latvian and translated to English) is harmful.

Appendix F: Regression Results

Table 5: OLS regression results: relationship between the dependent variable (harmful accepted proportion or harmless rejected proportion) and CommonCrawl corpus share (see appendix 3). observations on 24 eu official languages included in all cases. significance levels: *, **, *** represent 5%, 1%, and 0.1%, respectively.

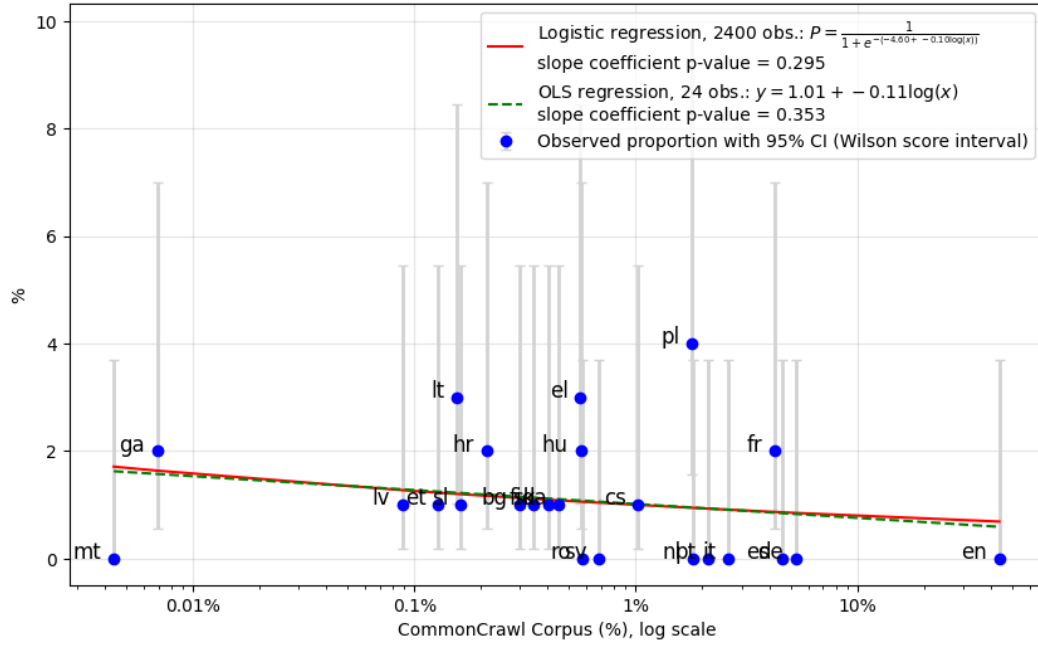| 2*Model	Harmful Accepted Proportion			Harmless Rejected Proportion		
	Intercept	Slope	p-value	Intercept	Slope	p-value
Claude 3.5 Sonnet	1.01	-0.11	0.353	70.99	-0.22	0.695
Gemini 1.5	1.32	-0.15	0.355	88.72	0.22	0.586
GPT-4o	6.44	-1.05**	0.002	37.40	0.71	0.112
Meta Llama 405B	4.47	0.23	0.249	55.18	0.76	0.488
Mistral Large 2	7.68	-1.59***	0.000	38.77	-1.04*	0.047

Figure 2: Claude 3.5 Sonnet, Harmful Accepted Proportion (100 observations per language)
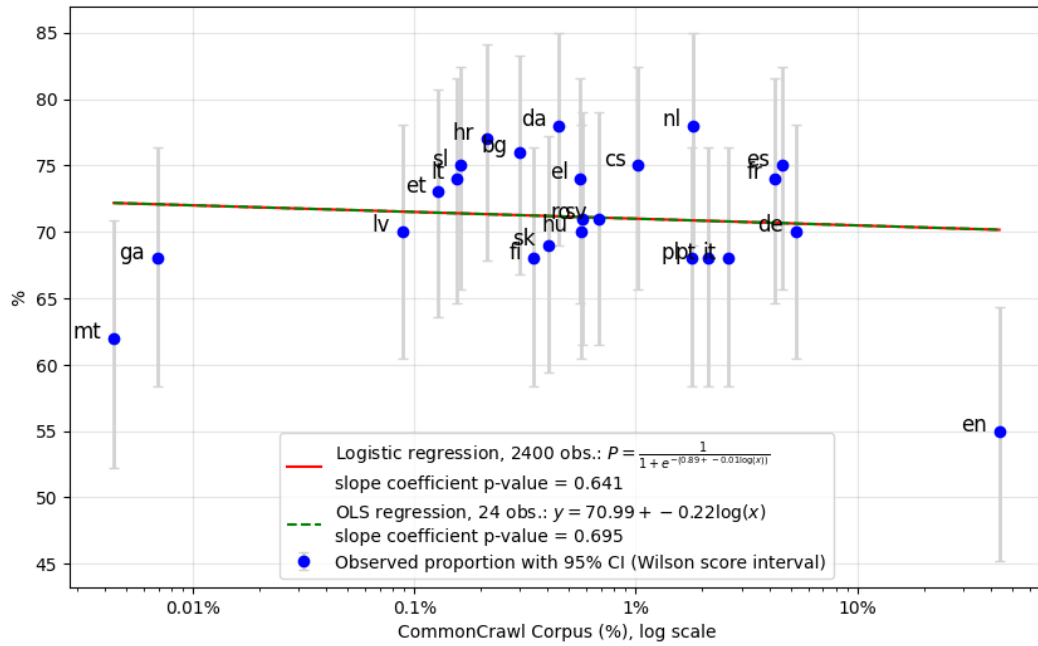

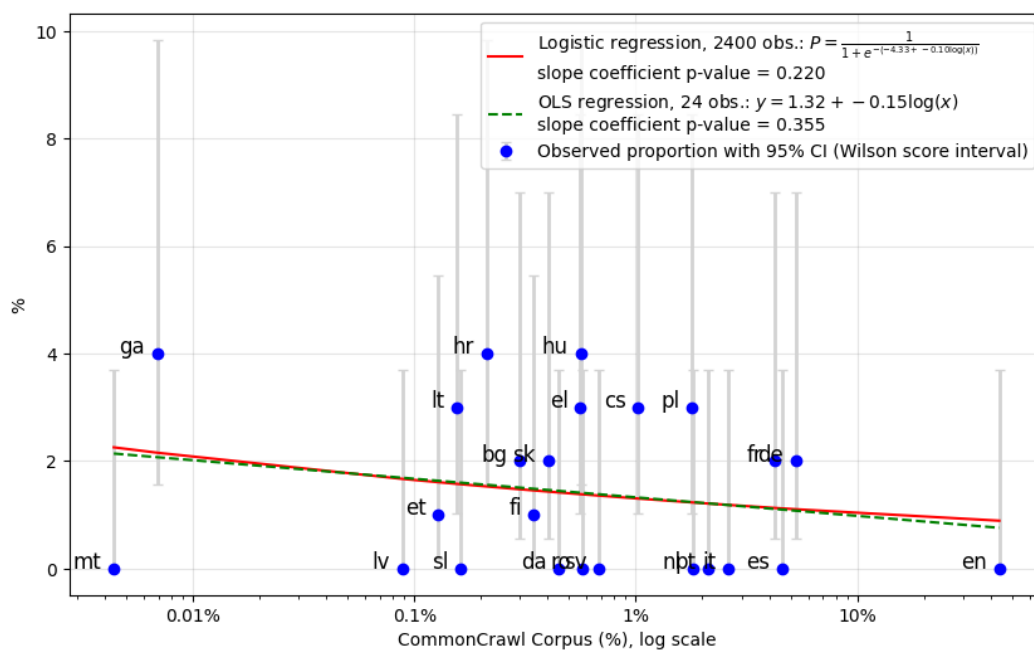Figure 3: Claude 3.5 Sonnet, Harmless Rejected Proportion (100 observations per language)

18

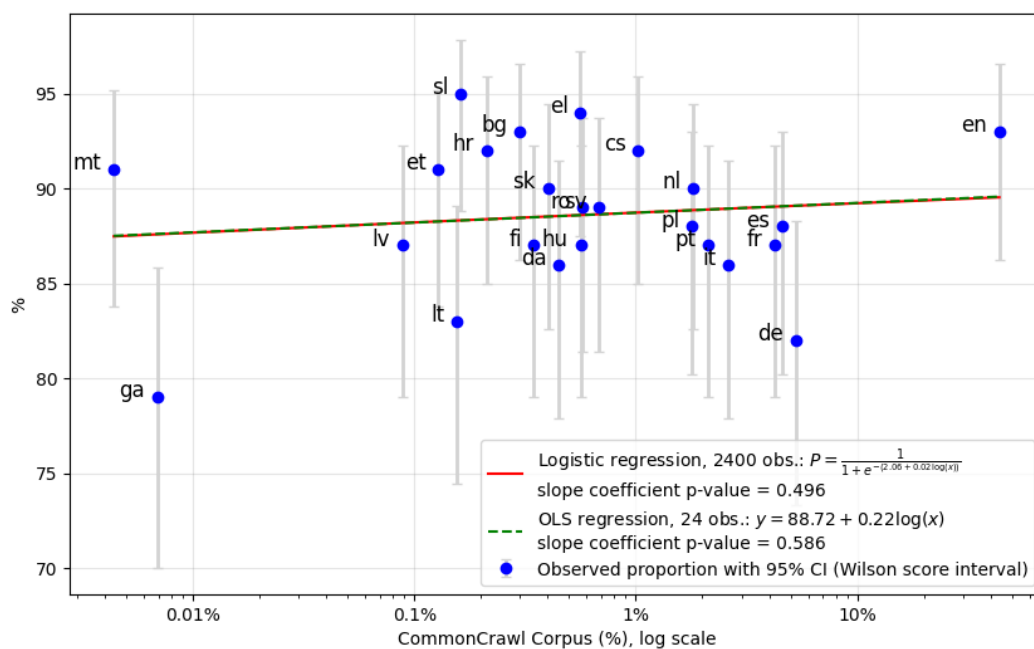Figure 4: Gemini 1.5 Pro, Harmful Accepted Proportion (100 observations per language)

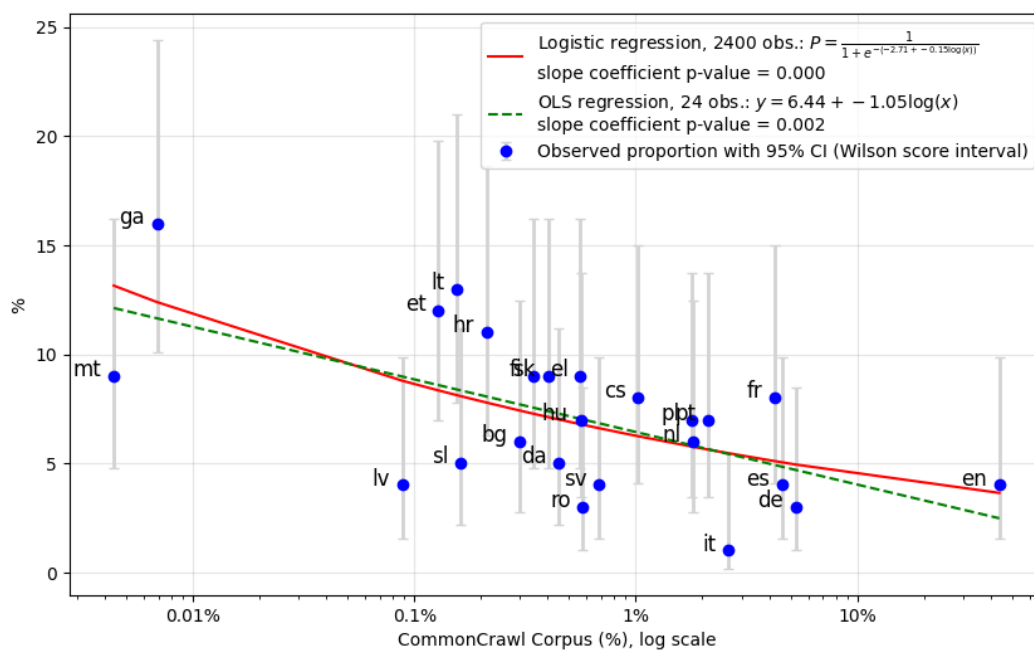Figure 5: Gemini 1.5 Pro, Harmless Rejected Proportion (100 observations per language)

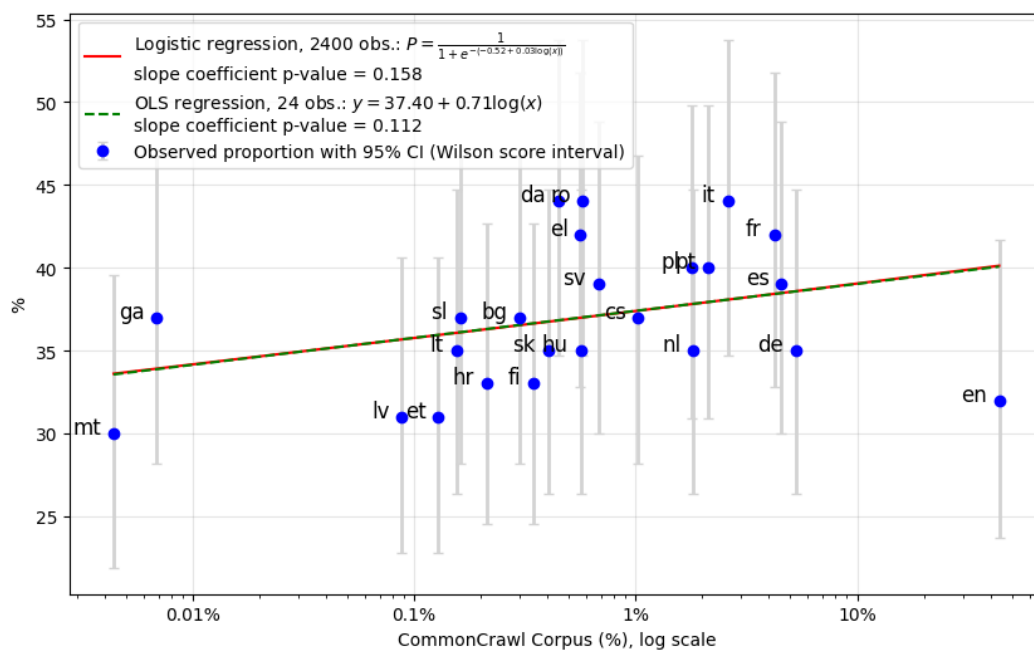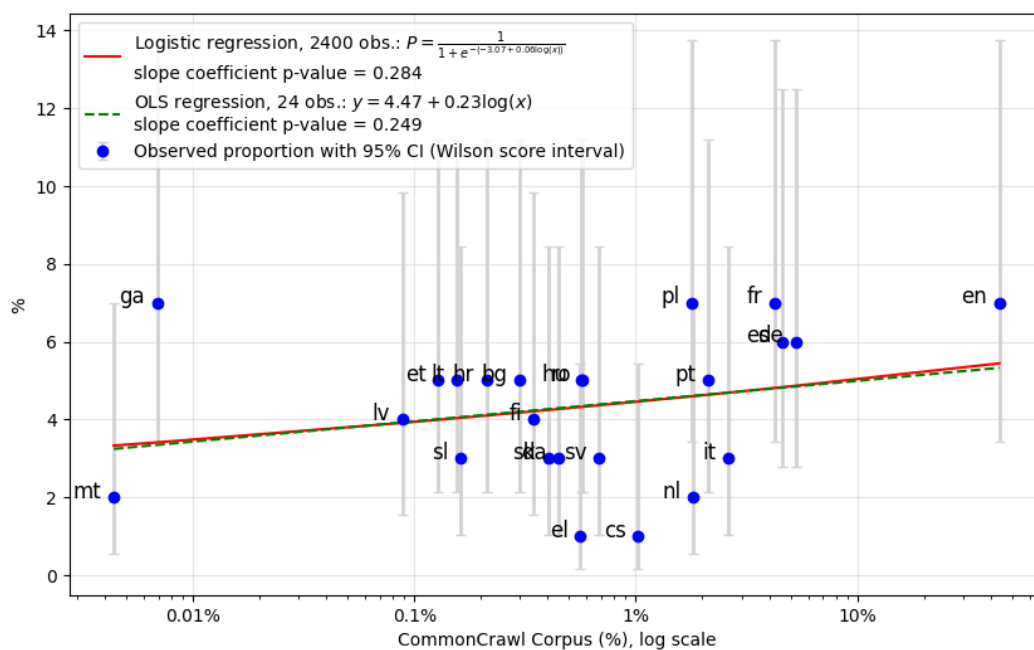Figure 6: GPT-4o, Harmful Accepted Proportion (100 observations per language)

Figure 7: GPT-4o, Harmless Rejected Proportion (100 observations per language)

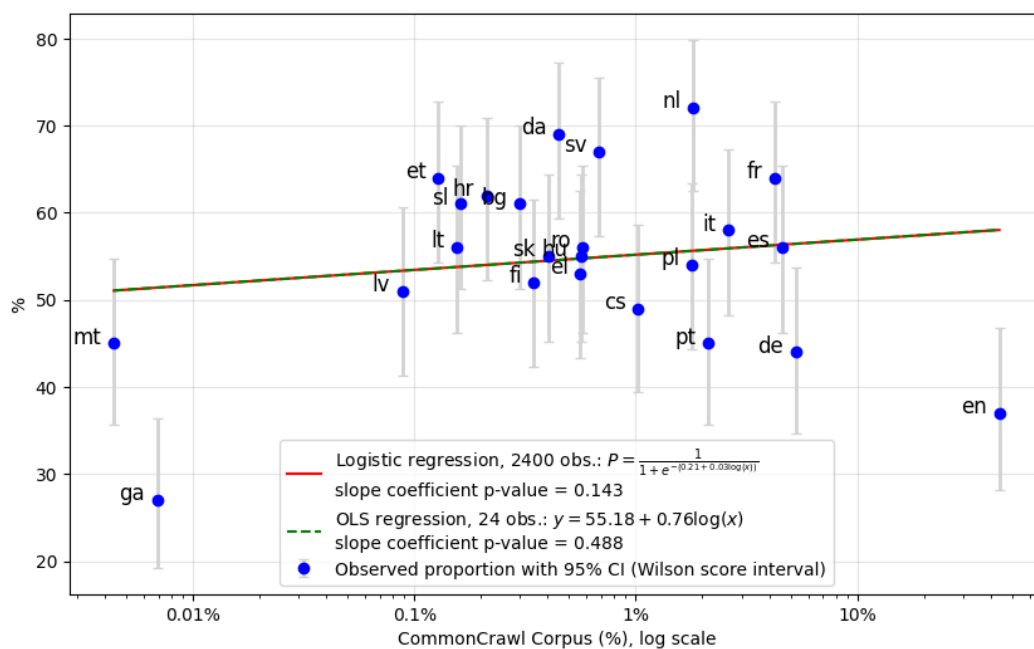Figure 8: Llama 3.1 405B, Harmful Accepted Proportion (100 observations per language)

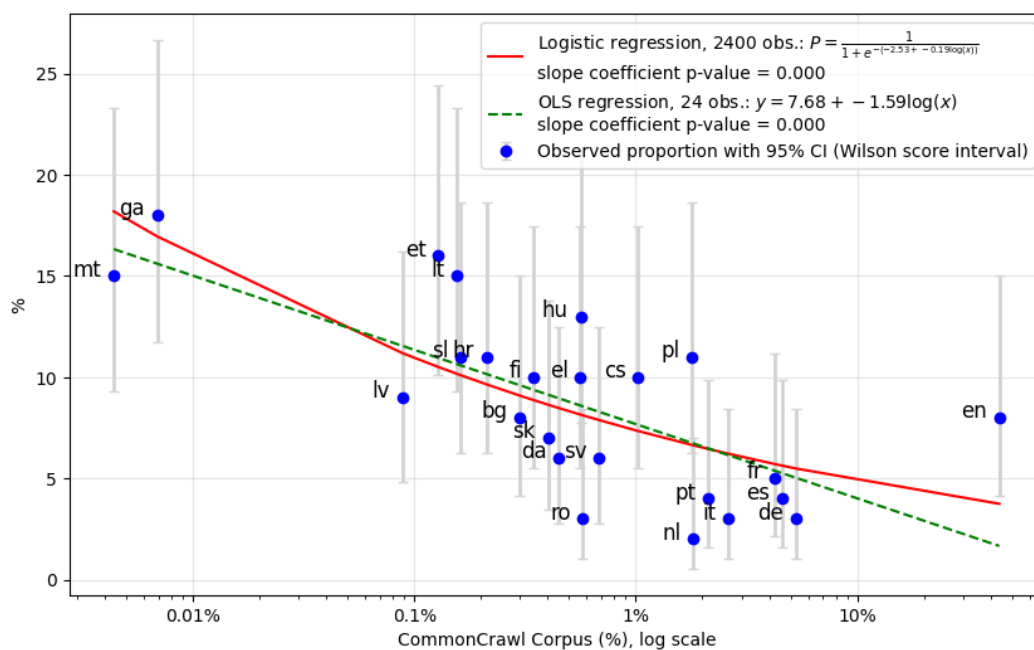Figure 9: Llama 3.1 405B, Harmless Rejected Proportion (100 observations per language)

Figure 10: Mistral Large 2, Harmful Accepted Proportion (100 observations per language)
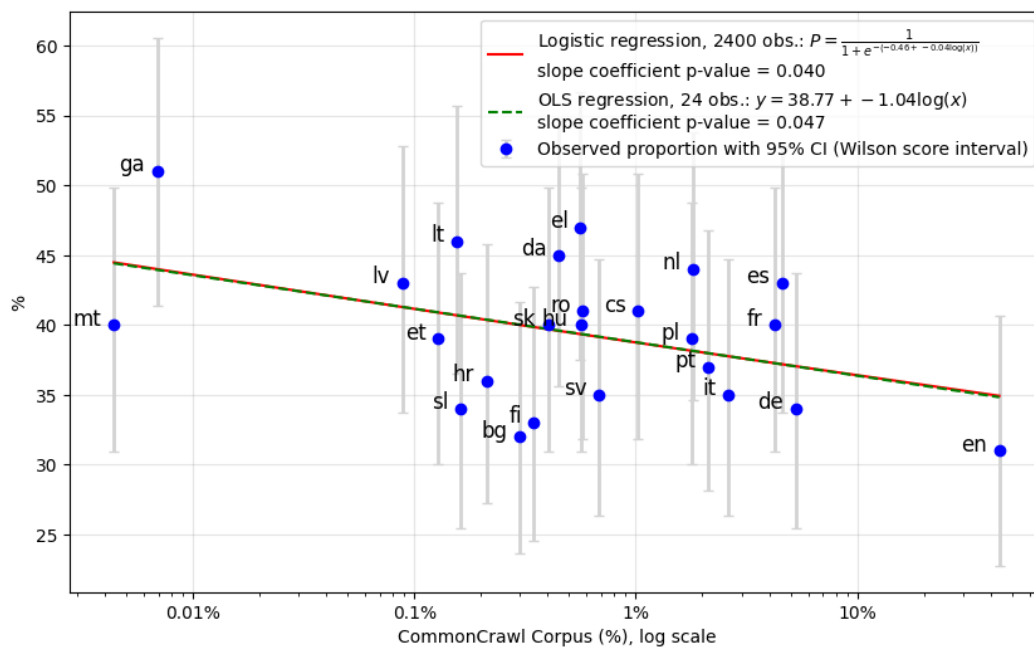

Figure 11: Mistral Large 2, Harmless Rejected Proportion (100 observations per language)