Safety in LLMs: Offensive Content, Cultural Region-Specific Sensitivity, Disinformation Project Proposal for NLP Course, Winter 2025

Dominik Zieliński
WUT
01171615@pw.edu.pl

Michał Korwek
WUT
01202291@pw.edu.pl

Ksawery Wojtaszek
WUT
01171419@pw.edu.pl

supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

1 Introduction

1.1 Project Proposal

The goal of our project is to develop a dedicated segment of the PW LLM Safety Bench focused on three high-impact risk categories: Offensive Content, Cultural & Region-Specific Sensitivity, and Disinformation. Based on our literature review, these areas represent core challenges in LLM safety, where current models still exhibit harmful behaviors such as toxic language generation, cultural bias, stereotyping, and the spread of false or misleading information.

Each prompt in the dataset will follow a unified specification to ensure consistency and reproducibility. For every entry, we will include: prompt ID, risk category, prompt text, safe/unsafe variant, expected model behavior, data type (text / image / mixed) and evaluation criteria. This structure provides a standardized format for testing and scoring model responses across all categories.

To evaluate LLM responses, we will apply a concise but effective set of safety metrics: Safety Compliance Rate - verifies whether the model appropriately refuses harmful or unsafe instructions. Toxicity Score - measures the presence of offensive or hateful language using tools such as Detoxify. Cultural Sensitivity Indicators - detect stereotyping, cultural bias, or region-specific misunderstandings. Factuality Score - evaluates correctness for disinformation-related prompts, particularly those involving common misconceptions. Hallucination Frequency - identifies fabricated or culturally inaccurate statements in model outputs. The evaluation will be performed using open-source LLMs, safety classifiers, and manual expert review of a subset of prompts. All generated data, prompt templates, and evaluation scripts will be stored in a reproducible format (Git repository) and accompanied by short documentation.

1.2 Tools & Models

For our benchmark, we will use a small but relevant set of established tools, open LLMs, and public datasets aligned with our three safety categories.

Tools: Detoxify - toxicity detection for offensive content. Perspective API - scoring hate, insult, profanity. HuggingFace Evaluate - factuality and toxicity evaluation.

Open Pre-Trained Models: LLaMA-3 8B, Mistral 7B, Gemma 7B, - representative open LLMs to test safety behavior. Qwen3 0.6B - the most lightweight model from qwen3 series Qwen3 8B - latest qwen model with number of parameters similar to llama, mistral and gemma for reliable comparison gpt-oss 20B - latest openai model with more parameters than other models selected

2 Literature Review

Large Language Models (LLMs) bring remarkable capabilities but also raise serious safety concerns in content generation. Key issues include the production of toxic or offensive content, lack of cultural or region-specific sensitivity, and the spread of disinformation or false information. Recent research in each of these areas has led to state-of-the-art models and datasets that aim to make LLMs safer and more trustworthy. Below, we highlight notable advances - targeting offensive/harmful content, region-specific sensitivity and addressing factual accuracy.

2.1 Mitigating Offensive Content with Constitutional AI

Bai et al. (Bai et al., 2022) propose Constitutional AI, a novel alignment framework that trains LLMs to be helpful yet harmless without relying on extensive human-labeled toxic data. Instead, the model is guided by a small set of explicit principles - a ”constitution” of AI-written rules - that it uses to critique and revise its own responses.

Through this two-stage process (illustrated in the figure), the model learns to refuse or filter offensive queries in a non-evasive manner. For example, when confronted with a harmful request, the model will politely explain its inability to comply rather than producing disallowed content or giving a generic refusal. The resulting tuned model (called RL-CAI) achieved state-of-the-art safety behavior: human evaluators preferred its answers over those from a baseline RLHF-trained model, finding it less harmful without sacrificing helpfulness. This work demonstrated that carefully crafted AI feedback and principles can significantly reduce toxic outputs while maintaining the model's usefulness, marking a key advance in offensive content mitigation.

2.2 Improving Truthfulness to Combat Disinformation

To address the risk of LLM-generated misinformation, Lin et al. (Lin et al., 2022) introduced TruthfulQA, a rigorous benchmark for measuring how truthfully models answer questions designed to elicit common misconceptions. This dataset consists of 817 diversified questions across domains like health, law, finance, and politics - queries that many humans answer incorrectly due to prevalent false beliefs. Evaluating several models on TruthfulQA yielded sobering results: even the best large model (GPT-3, 175B) answered only 58% of questions truthfully, whereas humans achieved 94% on the same set. Models frequently produced false but plausible-sounding answers - essentially mirroring popular myths or conspiracy theories - which could easily deceive users. Paradoxically, larger language models were less truthful on these adversarial questions than smaller ones, since bigger models more eagerly mimic the human text (and its misconceptions) found in their training data. This inverse scaling phenomenon highlights that simply making models bigger or training on more internet text will not solve the disinformation problem. Instead, the TruthfulQA study suggests that new fine-tuning strategies or objective functions are needed to improve truthfulness beyond human-imitated knowledge. By providing a standard benchmark for factual accuracy, TruthfulQA has spurred research into techniques (like retrieval augmentation and honesty-conditioned training) to ensure LLMs do not propagate false or misleading information.

Overall, these works exemplify the cutting-edge efforts to align LLM behavior with safety goals. Constitutional AI shows how a model can be trained to handle offensive content requests responsibly, and TruthfulQA exposes the gap in truthful reasoning that future models must bridge. Continued progress in such targeted subfields - from toxicity prevention to cultural sensitivity and truthfulness - is crucial for developing AI systems that are not only smart but also safe and respectful in a global context.

2.3 Cultural & Region-Specific Sensitivity in LLMs

As LLMs are deployed globally, a critical safety concern is their ability to understand cultural norms, avoid stereotyping, and generate region-appropriate responses. Even highly capable models often fail to recognize culturally sensitive topics or produce respectful, contextually aware answers across different regions or demographic groups. Below, we highlight two state-of-the-art research efforts that examine cultural sensitivity failures.

2.3.1 Measuring Cultural Biases and Stereotypes in LLMs

Parrish et al. (Parrish et al., 2022) introduced the BBQ dataset (Bias Benchmark for Question Answering), one of the largest and most influential benchmarks for evaluating social, cultural, and demographic bias in language models. The dataset contains over 58,000 question-answer pairs targeting bias across categories such as nationality, ethnicity, religion, disability, gender identity, and socio-economic status. Each question embeds subtle cultural context, allowing the benchmark to detect whether models rely on stereotypes instead of factual reasoning.

Their evaluation shows that major LLMs systematically favor stereotypical answers when the question is ambiguous or under-specified. For instance, models frequently infer nationality or religion based purely on names, or associate certain ethnic groups with negative attributes. Importantly, these biases persist even in very large models trained on web-scale datasets, suggesting that cultural stereotypes are deeply embedded in the data. BBQ has since become a standard reference for auditing cultural and socio-regional bias in LLMs.

2.3.2 LLM in the context of different cultures

As LLMs are increasingly deployed in non-English contexts, ensuring their safety requires understanding local cultural norms, legal frameworks, and societal sensitivities.For example Li et al. (Li et al., 2025) introduce LiveSecBench, a dynamic benchmark designed specifically for evaluating LLM safety in the Chinese linguistic and cultural context. Unlike general-purpose or English-centric safety datasets, LiveSecBench incorporates prompts that reflect local laws, ethical considerations, privacy concerns, and region-specific adversarial risks.

2.3.3 Bilingual Context of LLM

Using LLM's is particularly interesting in bilingual countries such as Kazakhstan, where both Russian and Kazakh are used on a daily basis. Countries with such traits may create an environment where two languages are mixed, to the degree some words in one sentence come from one language and some come from the other. Goloburda et al. (Goloburda et al., 2025) saw this fact and decided to check how LLM works in this bilingual context.

Their study demonstrates that mixed-language prompts can weaken or bypass standard safety mechanisms present in LLMs. As a result, models may produce content that would normally be filtered out in monolingual settings. For example a prompt written partly in Russian and partly in Kazakh asking about a controversial political event produced a confident but incorrect explanation - a form of disinformation that the same model did not generate when the prompt was given entirely in Russian.

The authors show that models often fail to recognise culturally sensitive topics specific to Central Asia, such as ethnic relations or political tension, leading to biased or inappropriate responses. Furthermore, limited training data in Kazakh makes LLMs more susceptible to generating or amplifying misleading narratives, especially in politically charged contexts.

2.3.4 Cultural Sensitivity Failures in Open-Domain LLMs

Recent work shows that text-only LLMs often struggle with culturally grounded reasoning. Li et al. (Li et al., 2024) introduce CultureLLM, demonstrating that mainstream LLMs frequently reflect Western-centric opinion distributions because English-language data dominate pre-training. Using only 50 culturally sensitive seed questions from the World Values Survey and a semantic data augmentation pipeline, the authors fine-tune culture-specific and unified LLMs that better capture regional values. Across tasks such as hate speech detection, bias detection, toxicity classification, and stance detection, CultureLLM substantially outperforms GPT-3.5 and Gemini Pro, including for low-resource cultures, showing that lightweight cultural fine-tuning improves alignment. Beyond text-only models, similar gaps emerge in multimodal settings. Nayak et al. (Nayak et al., 2024) conduct a large-scale study of cultural understanding in VLMs using the CULTURALVQA benchmark, which evaluates models across cultural domains such as food, drinks, clothing, rituals, and traditions. Their results show cross-regional disparities in performance. State-of-the-art systems such as GPT-4 achieve strong accuracy on North American cultural concepts (up to 72%) but drop sharply for African-Islamic regions (as low as 43%). Open-source models perform even worse, with the best model (InternVL) averaging only 46% accuracy. The authors also show that VLMs struggle with country-specific cultural knowledge, even when the visual information is clear, frequently misidentifying culturally relevant objects, confusing similar items, or failing to capture the cultural significance of practices and symbols. Overall, research on cultural and region-specific sensitivity reveals that LLMs often fail to adapt their behavior to diverse cultural norms, leading to stereotypes, misinformation, and insensitive responses. Datasets such as BBQ and CULTURALVQA provide structured ways to measure these failures and highlight the need for culturally aware alignment strategies. Ensuring that LLMs behave respectfully and accurately across regions is a key component of LLM safety and is essential for global deployment.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson

Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback.

Maiya Goloburda, Nurkhan Laiyk, Diana Turmakhan, Yuxia Wang, Mukhammed Togmanov, Jonibek Mansurov, Askhat Sametov, Nurdaulet Mukhituly, Minghan Wang, Daniil Orel, Zain Muhammad Mujahid, Fajri Koto, Timothy Baldwin, and Preslav Nakov. 2025. Qorgau: Evaluating llm safety in kazakh-russian bilingual contexts.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models.

Yudong Li, Zhongliang Yang, Kejiang Chen, Wenxuan Wang, Tianxin Zhang, Sifang Wan, Kecheng Wang, Haitian Li, Xu Wang, Lefan Cheng, Youdan Yang, Baocheng Chen, Ziyu Liu, Yufei Sun, Liyan Wu, Wenya Wen, Xingchi Gu, and Peiru Yang. 2025. Livesecbench: A dynamic and culturally-relevant ai safety benchmark for llms in chinese context.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. Bbq: A hand-built bias benchmark for question answering.