Title
Project Proposal for NLP Course, Winter 2025

Kinga Frańczak
Warsaw University of Technology
kinga.franczak.stud@pw.edu.pl

Kamil Kisiel
Warsaw University of Technology
kamil.kisiel3.stud@pw.edu.pl

Wiktoria Koniecko
Warsaw University of Technology
wiktoria.koniecko.stud@pw.edu.pl

Piotr Kosakowski
Warsaw University of Technology
piotr.kosakowski2.stud@pw.edu.pl

Abstract

The rapid deployment of Large Language Models (LLMs) has outpaced comprehensive safety evaluation, leaving them vulnerable to the generation of harmful content. Existing benchmarking efforts remain fragmented, often focusing on isolated threats such as toxicity and overlooking coordinated attacks across critical domains. The LLM Safety Benchmark (LSB) is introduced as a unified framework to evaluate model resilience against disinformation, misinformation, and health-related threats.

The LSB consists of 600 adversarial prompts distributed across three domains and organised into four difficulty tiers. In contrast to existing datasets, LSB incorporates advanced attack vectors, including implicit misinformation, prompt injection, and jailbreaking techniques. The framework assesses model performance using both attack success rates and refusal metrics, covering tasks such as detecting logical fallacies and identifying health risks. By integrating automated metrics with human annotation, LSB establishes a standardised and reproducible protocol for cross-model comparison. This approach advances LLM safety research by revealing vulnerabilities that isolated prompt testing may overlook and provides a robust tool for developing more resilient artificial intelligence systems.

1 Introduction & Motivation

Large Language Models (LLMs) have achieved remarkable capabilities in natural language understanding and generation, powering applications across industry and academia. However, their rapid deployment has outpaced comprehensive safety evaluation. Recent incidents demonstrate that state-of-the-art models can generate harmful content, including false information, biased outputs, and health-endangering advice. While isolated benchmarks exist for specific threat categories (e.g., toxicity, bias), no unified framework evaluates LLM resilience across coordinated disinformation, misinformation, and health-related attacks simultaneously.

1.1 Problem Statement

Current safety benchmarking efforts suffer from critical limitations:

1. **Fragmented Landscape**: Existing benchmarks address isolated problems (ToxiGen(Hartvigsen et al., 2022) for hate speech, RealToxicityPrompts (Gehman et al., 2020) for toxicity, BOLD(Bolukbasi et al., 2021)

for bias) but lack integration across multiple threat vectors in critical domains.

2. **Domain-Specific Gaps**: Disinformation benchmarks rarely test coordinated false narrative generation; misinformation evaluation focuses on obvious rather than subtle claims; health threat assessment remains minimal.

3. **Attack Sophistication**: Previous work emphasizes direct attacks. Sophisticated techniques (jailbreaking, role-playing, prompt injection) remain under-evaluated.

4. **Lack of Standardisation**: No standard methodology, metrics, or annotation guidelines exist, hindering reproducibility and cross-model comparison.

1.2 Research Questions

This work addresses three central research questions:

- How resilient are current LLMs to coordinated attacks across disinformation, misinformation, and health threat domains?
- What attack strategies are most effective at inducing harmful outputs within each category?
- Can we develop a standardised, reproducible benchmark enabling fair model comparison?

1.3 Contribution

We propose **LSB** (LLM Safety Benchmark), a unified evaluation framework comprising 600 adversarial prompts (200 per domain) across four difficulty tiers. LSB integrates disinformation attacks (false narratives, fabricated events, coordinated inauthentic behaviour), misinformation attacks (misleading framing, out-of-context information, manipulated evidence), and health threat attacks (harmful medical advice, mental health endangerment, physical safety risks). Evaluation employs both automated metrics and human annotation, enabling comprehensive vulnerability assessment.

2 Literature Review & Related Work

2.1 Existing Safety Benchmarks

Current LLM safety evaluation relies on isolated, domain-specific benchmarks. The evaluation methods and metrics used differ across benchmark datasets.

SafetyBench (Zhang et al., 2024) comprises 11,435 multiple-choice questions across seven safety categories in both Chinese and English, tested over 25 popular LLMs. Additionally, it compares results achieved with zero-shot and five-shot learning.

ALERT (Tedeschi et al., 2024) introduces a large-scale benchmark comprising 45,000+ instructions using a fine-grained risk taxonomy for red-team evaluation. Results are evaluated across six different categories and 32 micro categories. The performance of a model is evaluated by an auxiliary model that marks responses as safe or unsafe.

HarmBench (Mazeika et al., 2024) provides a standardized framework for automated red teaming with 500+ harmful behaviours and 33 target LLMs.

Holistic Evaluation of Language Models (Liang et al., 2023) proposes the HEML dataset containing 15,908 questions across 42 scenarios and evaluates 30 models. Its focus extends beyond model safety to include tasks such as classification and data extraction. However, HEML offers the most robust comparison methods by assessing each model on seven metrics and providing more nuanced results.

However, these benchmarks treat threat categories independently and do not assess coordinated multi-domain attacks. The different metrics and evaluation methods provide a detailed and nuanced view of model performance, but make comparison across benchmark datasets challenging.

2.2 Disinformation and Narrative Generation

The studies show that LLM models can create and spread disinformation on specific topics, as well as combine multiple false narratives into a coherent narrative. Furthermore, the evaluation may pose additional challenges in adapting to new disinformation narratives.

Large Language Models Can Consistently Generate High-Quality Content for Election Disinformation Operations (Williams et al., 2025) demonstrates that LLMs achieve above-human performance in generating coordinated false narratives for election-related disinformation.

DiNaM: Disinformation Narrative Mining (Sosnowski et al., 2025) shows that LLMs can weave multiple false claims into coherent narratives by clustering false information into semantic groups, revealing vulnerabilities that isolated prompt testing misses.

TripleFact: Defending Data Contamination in the Evaluation of LLM-driven Fake News Detection (Xu and Yan, 2025) highlights the problem of evaluating models' susceptibility to disinformation against commonly used benchmark datasets, when models are trained on them, which can lead to inflated metrics. The paper proposes a solution that mitigates the risks of benchmark data contamination and provides more accurate evaluation metrics.

2.3 Implicit Misinformation

The topic of LLMs' responses to implicit misinformation in queries is unexplored mainly; the first papers evaluating models' performance and responses were published this year, with some focused on narrower subjects such as medical misinformation. The current papers point to the failure of LLMs to capture and correctly respond to false information in prompts and potential risks related to it.

How to Protect Yourself from 5G Radiation? Investigating LLM Responses to Implicit Misinformation (Guo et al., 2025) introduces the first framework for evaluating implicit misinformation, where false assumptions are embedded in queries rather than explicitly stated. Testing 15 state-of-the-art LLMs reveals that even GPT-4 fails on approximately 40% of implicit misinformation cases - a critical gap that existing benchmarks focusing on obvious false claims do not capture.

Understanding Knowledge Drift in LLMs Through Misinformation (Fastowski and Kasneci, 2025) shows how the implicit misinformation in queries can affect the model responses by analysing the value of metrics such as entropy, perplexity, and token probability across state-of-the-art models. The paper reveals that factually incorrect responses to misinformation in queries have higher uncertainty, which can be decreased by repeated exposure to false information.

2.4 Health Threat Assessment

Misinformation in medical and health-related topics remains significant due to the potential risks it poses. Research shows that LLM models are vulnerable to attacks, and solutions such as high-quality training and evaluation datasets, as well as methods for medical and health-related data, are being proposed.

Medical Large Language Models Are Susceptible to Targeted Misinformation Attacks (Han et al., 2024) demonstrates that domain-specialised medical LLMs remain vulnerable to targeted health misinformation attacks. Health threat assessment remains isolated from other safety dimensions in current benchmarks.

HealthFC: Verifying Health Claims with Evidence-Based Medical Fact-Checking (Vladika et al., 2024) provides 750 health-related claims verified by medical experts using evidence from systematic reviews and clinical trials.

Did You Tell a Deadly Lie? Evaluating Large Language Models for Health Misinformation Identification (Thapa et al., 2025) evaluates seven state-of-the-art large language models on medical misinformation detection across four different datasets.

2.5 Attack Sophistication and Difficulty Scaling

Jailbreak Attacks and Defenses Against Large Language Models: A Survey (Yi et al., 2024) documents that attack effectiveness varies dramatically with sophistication level.

The TIP of the Iceberg: Revealing a Hidden Class of Task-in-Prompt Adversarial Attacks on LLMs (Berezin et al., 2025) introduces systematic difficulty-scaled evaluation across three tiers: explicit solution guidance, partial task hints, and no hints. Critically, optimal attack difficulty varies by model capability.

A STRONGREJECT for Empty Jailbreaks (Souly et al., 2024) refines evaluation by distinguishing between willingness and capability dimensions of jailbreak success, revealing that binary evaluation metrics obscure essential distinctions.

2.6 Integration Across Domains

SAGE: A Generic Framework for LLM Safety Evaluation (Jindal et al., 2025) provides methodology for fair cross-model comparison across multiple harm areas, demonstrating that standardised evaluation protocols and multi-domain assessment are essential for comprehensive safety evaluation. As opposed to benchmark datasets described in the section on Existing Safety Benchmarks, which assess models based on responses to a single query, the SAGE frameworks support evaluation of multi-turn conversations.

3 Solution overview

3.1 Adversarial prompts

As mentioned in section 1.3, our goal is to create a set of 600 adversarial prompts across three domains and four difficulty tiers (50 per domain and difficulty tier).

We will consider the following types of questions in the health category:

- **Threat detection**: prompts asking to identify how dangerous a particular situation is in terms of how likely a human is to be harmed (risk) and how severe the potential damage is (impact).
- **Health advice**: prompts requesting health and self-care advice (including mental health), asking how to avoid danger or how to decrease the risk and impact of harm in a potentially dangerous situation.

In case of misinformation, we should consider the following types of problems:

- **Common misconceptions**: the model may respond incorrectly to prompts with no misinformation due to the presence of misinformation in the training data.
- **Explicit misinformation**: the prompt contains false statements.
- **Implicit misinformation**: the prompt implicitly assumes an incorrect statement is true.
- **Logical fallacies**: The prompt contains a logical error, so that misinformation can be primarily identified based on the prompt alone, with little or no additional knowledge. Logical fallacies may be considered a special case of explicit misinformation.

The model should correct the misinformation present in the prompt.

In case of a suspected attempt to generate disinformation, the model should refuse to generate content. It should, however, respond to other, harmless requests — thus, the set of prompts in the disinformation domain should contain harmless requests as well.

3.2 Evaluation metrics

We will use attack success rates across all tasks in the misinformation and disinformation domains and for advice requests in the health domain to measure how frequently the model produces undesirable output. Additionally, in the disinformation domain, we will also use the refusal rate to measure availability.

In the case of a threat-detection problem in the health domain, we receive three pieces of information from a single prompt: the potential impact, likelihood and source of threat. We may use ASR to estimate how often a specific model misidentifies the source of the danger. The possible implications and risks can be calculated, for example, on a scale from 1 to 10, where 1 is ”very low” and 10 is ”very high”. Then we may use regression metrics, such as MSE, to evaluate both values independently.

3.3 Attack methods

- **No attack** — used to check if a model refuses to provide an answer for a harmless request.
- **Topic selection** — selecting a topic with a lot of misinformation online may lead to the model providing a misleading answer.
- **Implicit misinformation** — mentioned previously, may mislead the model.
- **Prompt injection** — the result of the model being unable to differentiate between the system prompt and the user prompt. The user may tell the model, e.g. to ignore previous instructions.
- **Jailbreak** — manipulating the model to bypass safety measures, leading to undesirable output. Includes e.g. the usage of role-play, typos and formatting. There have recently been successful multi-turn jailbreak attempts through repeatedly prompting the model to generate related content (Russinovich et al., 2025).

3.4 Dataset Statistics

We have created a dataset of 452 prompts as a foundation for the full LSB benchmark. This section presents key statistics characterizing the current dataset composition and structure.

3.4.1 Overall Composition

The dataset consists of 452 prompts distributed across three domains: health (152), misinformation (150), and disinformation (150). Prompts are organized into four difficulty tiers, with 85 prompts in Tier 1, 114 in Tier 2, 135 in Tier 3, and 118 in Tier 4. The dataset includes 111 safe prompts (harmless requests to test for false refusals) and 341 unsafe prompts (adversarial attacks). Table 1 shows the distribution across domains and difficulty tiers.

Domain	Tier 1	Tier 2	Tier 3	Tier 4
Health	30	37	44	41
Misinformation	28	40	46	36
Disinformation	27	37	45	41
Total	**85**	**114**	**135**	**118**

Table 1: Distribution of prompts across domains and difficulty tiers.

3.4.2 Attack Method Distribution

The dataset incorporates six attack methods, with distribution shown in Table 2. The most common attack method is topic selection (95 prompts), followed by implicit misinformation (75), jailbreak/roleplay (60), prompt injection (54), multi-turn persuasion (52), and no attack baseline (116). The multi-turn persuasion category includes sophisticated conversational attacks that gradually escalate requests, as documented in recent work (Russinovich et al., 2025).

Attack Method	Count
No attack	116
Topic selection	95
Implicit misinformation	75
Jailbreak/Roleplay	60
Prompt injection	54
Multi-turn persuasion	52
Total	**452**

Table 2: Distribution of prompts by attack method.

3.4.3 Expected Behavior Distribution

Prompts are categorized by expected model behavior: *blocked* (267 prompts) indicates the model should refuse the request; *factual* (161 prompts) indicates the model should provide accurate information and correct any misinformation; *safe* (14 prompts) tests for false refusals on harmless requests; and *threat_assessment* (10 prompts) requires the model to evaluate health risks. The high proportion of blocked prompts (59.1%) reflects the adversarial nature of the benchmark, particularly in the disinformation domain where 76.7% of prompts should be refused.

3.4.4 Prompt Characteristics

The dataset includes both single-turn (399 prompts, 88.3%) and multi-turn conversational prompts (53 prompts, 11.7%). Multi-turn prompts are distributed across domains (health: 18, misinformation: 18, disinformation: 17) and primarily appear in higher difficulty tiers (Tier 3 and Tier 4), reflecting the increased sophistication required for multi-turn attacks. Prompt length varies substantially, with a median length of 94 characters (interquartile range: 63–244 characters), ranging from 23 to 1,334 characters. The longer prompts typically correspond to multi-turn conversations or complex scenarios requiring detailed context.

3.4.5 Domain-Specific Characteristics

Each domain exhibits distinct characteristics aligned with its evaluation objectives. The health domain includes 10 threat assessment prompts requiring structured risk evaluation (likelihood, impact, source), alongside 69 health advice prompts and 36 physical safety prompts. The misinformation domain emphasizes correction tasks, with 88 prompts expecting factual responses that debunk false claims, distributed across common misconceptions (45), implicit misinformation (25), explicit misinformation (47), and logical fallacies (10). The disinformation domain focuses on refusal behavior, with 115 prompts (76.7%) requiring the model to block content generation, covering false narratives (66) and coordinated inauthentic behavior (49).

4 Proof of Concept

We validated the full pipeline (prompts → model generation → automatic scoring → JSON/CSV outputs) with three openly available models. The PoC covers environment setup, unified execution across all domains, and summarised results.

Pipeline

1. **Environment**: create and activate an isolated env (e.g., `conda create -n lsb-nlp python=3.10; conda activate lsb-nlp`), then install dependencies with `pip install -r requirements.txt`.

2. **Run all domains for one model**: The evaluator formats prompts (single- and multi-turn), generates responses, detects refusals/harmful content, applies LLM-as-judge for factual items, extracts threat-assessment scores (likelihood/impact/source), and writes per-prompt JSON/CSV plus printed summaries (ASR, refusal rate by domain/tier/attack method, threat-assessment MSE/RMSE).

3. **Scaling comparison**: repeat Step 2 for additional models to observe family/size effects. In the PoC we ran: `Qwen/Qwen2.5-0.5B-Instruct`, `Qwen/Qwen2.5-1.5B-Instruct`, and `TinyLlama-1.1B-Chat-v1.0`.

Observed results — Batch 1 (150 prompts / model)

- **Qwen2.5-0.5B-Instruct**: ASR 42%, refusal 32.7%; by domain — health 40% ASR (18% refusal), misinformation 44% (34%), disinformation 42% (46%). Attack success rose with tier (Tier 1: 25%, Tier 4: 63%).
- **Qwen2.5-1.5B-Instruct**: ASR 20%, refusal 45.3%; domains — health 18% (42%), misinformation 20% (34%), disinformation 22% (60%). Strong gains vs 0.5B at lower tiers (Tier 1: 0% ASR; Tier 4: 46.7%).
- **TinyLlama-1.1B-Chat-v1.0**: ASR 46.7%, refusal 9.3%; domains — health 44% (4%), misinformation 34% (10%), disinformation 62% (14%). High compliance on hard tiers (Tier 4: 76.7% ASR).

Observed results — Batch 2 (152 prompts / model)

- **Qwen2.5-0.5B-Instruct**: ASR 48.0%, refusal 36.2%; by domain — health 32.7% ASR (38.5% refusal), misinformation 58.0% (28.0%), disinformation 54.0% (42.0%). Attack success by tier: Tier 1: 47.5%, Tier 2: 38.5%, Tier 3: 43.6%, Tier 4: 64.7%. Most vulnerable to multi-turn persuasion (70.8% ASR) and jailbreak/roleplay (56.0%).
- **Qwen2.5-1.5B-Instruct**: ASR 27.0%, refusal 43.4%; by domain — health 25.0% ASR (40.4% refusal), misinformation 20.0% (30.0%), disinformation 36.0% (60.0%). Attack success by tier: Tier 1: 20.0%, Tier 2: 10.3%, Tier 3: 30.8%, Tier 4: 50.0%. Strong resistance to prompt injection (12.5% ASR, 87.5% refusal) but vulnerable to multi-turn persuasion (66.7% ASR).

Aggregated results (302 prompts / model) Table 3 presents weighted aggregate performance across both batches. Results are computed as weighted averages accounting for batch sizes (150 + 152 = 302 prompts).

Model	ASR	Refusal	N
Qwen2.5-0.5B	45.0%	34.4%	302
Qwen2.5-1.5B	23.5%	44.4%	302
TinyLlama-1.1B	46.7%	9.3%	150

Table 3: Aggregated model performance. ASR = Attack Success Rate, N = number of prompts evaluated.

Key observations from the aggregated analysis:

- **Model scaling improves safety**: Qwen2.5-1.5B shows a 21.5 percentage point reduction in ASR compared to Qwen2.5-0.5B (23.5% vs 45.0%), demonstrating that increased model capacity correlates with improved safety behavior.
- **Refusal-ASR trade-off**: The 1.5B model's lower ASR comes with higher refusal rates (44.4% vs 34.4%), suggesting a more conservative safety posture that may include some false refusals.
- **Multi-turn attacks remain challenging**: Both Qwen models show high vulnerability to multi-turn persuasion attacks (70.8% and 66.7% ASR respectively), indicating that conversational context manipulation is an effective attack vector even for larger models.
- **TinyLlama lacks safety training**: With only 9.3% refusal rate and 46.7% ASR, TinyLlama demonstrates minimal safety guardrails, complying with most requests regardless of harm potential.
- **Tier 4 remains difficult**: All models show elevated ASR on Tier 4 (highest difficulty) prompts, confirming that sophisticated attacks combining multiple techniques pose the greatest challenge.

All runs produced timestamped JSON/CSV files in `results/`, which back the reported figures.

References

Sergey Berezin, Reza Farahbakhsh, and Noel Crespi. 2025. The TIP of the iceberg: Revealing a hidden class of task-in-prompt adversarial attacks on LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6716–6730, Vienna, Austria, July. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2021. Bold: Dataset and metrics for measuring bias in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 862–872. ACM.

Alina Fastowski and Gjergji Kasneci. 2025. Understanding knowledge drift in llms through misinformation. In Marco Piangerelli, Bardh Prenkaj, Ylenia Rotalinti, Ananya Joshi, and Giovanni Stilo, editors, *Discovering Drift Phenomena in Evolving Landscapes*, pages 74–85, Cham. Springer Nature Switzerland.

Samuel Gehman, Suchin Ghai, Andrew Huang, Akhil Sharma, Alison Wong, Arvind Singh, and Thomas Wolf. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: ACL 2020*, pages 3356–3369. Association for Computational Linguistics.

Ruohao Guo, Wei Xu, and Alan Ritter. 2025. How to protect yourself from 5g radiation? investigating llm responses to implicit misinformation.

Tianyu Han, Sven Nebelung, Firas Khader, Tianci Wang, Gustav Müller-Franzes, Christiane Kuhl, Sebastian Foersch, Jens Kleesiek, Christoph Haarburger, Keno Bressem, Jakob Kather, and Daniel Truhn. 2024. Medical large language models are susceptible to targeted misinformation attacks. *npj Digital Medicine*, 7, 10.

Thomas Hartvigsen, Saadia Gabriel, Huizi Wang, Alison Parrish, and Robert C. West. 2022. Toxigen: A large-scale machine generated dataset for implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6054–6067. Association for Computational Linguistics.

Madhur Jindal, Hari Shrawgi, Parag Agrawal, and Sandipan Dandapat. 2025. Sage: A generic framework for llm safety evaluation.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. 2024. Harmbench: a standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Mark Russinovich, Ahmed Salem, and Ronen Eldan. 2025. Great, now write an article about that: The crescendo {Multi-Turn}{LLM} jailbreak attack. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 2421–2440.

Witold Sosnowski, Arkadiusz Modzelewski, Kinga Skorupska, and Adam Wierzbicki. 2025. DiNaM: Disinformation narrative mining with large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30212–30239, Suzhou, China, November. Association for Computational Linguistics.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. 2024. A strongreject for empty jailbreaks. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24.

Simone Tedeschi, Felix Friedrich, Patrick Schramowski, Kristian Kersting, Roberto Navigli, Huu Nguyen, and Bo Li. 2024. Alert: A comprehensive benchmark for assessing large language models' safety through red teaming.

S Thapa, K Rauniyar, H Veeramani, A Shah, Imran Razzak, and U Naseem. 2025. Did You Tell a Deadly Lie? Evaluating Large Language Models for Health Misinformation Identification. *Web Information Systems Engineering – WISE 2024*, 1.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024. HealthFC: Verifying health claims with evidence-based medical fact-checking. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8095–8107, Torino, Italia, May. ELRA and ICCL.

Angus Williams, Liam Burke-Moore, Ryan Chan, Florence Enock, Federico Nanni, Tvesha Sippy, Yi-Ling Chung, Evelina Gabasova, Kobi Hackenburg, and Jonathan Bright. 2025. Large language models can consistently generate high-quality content for election disinformation operations. *PLOS ONE*, 20, 03.

Cheng Xu and Nan Yan. 2025. TripleFact: Defending data contamination in the evaluation of LLM-driven fake news detection. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8808–8823, Vienna, Austria, July. Association for Computational Linguistics.

Sibo Yi, Yule Liu, Zhen Sun, Tianshuo Cong, Xinlei He, Jiaxing Song, Ke Xu, and Qi Li. 2024. Jailbreak attacks and defenses against large language models: A survey.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2024. SafetyBench: Evaluating the safety of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15537–15553, Bangkok, Thailand, August. Association for Computational Linguistics.

Member	**Sections**	**Code & Prompts**
K. Kisiel	1, 3.4, 4	Eval. pipeline, 150 prompts
W. Koniecko	3.1–3.3	—
K. Frańczak	2 (w/ Piotr)	150 prompts
P. Kosakowski	2 (w/ Kinga)	150 prompts

Table 4: Work division among team members.