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

This section describes the comprehensive evaluation framework developed for the LSB benchmark, including the adversarial prompt design, evaluation methodology, attack methods, and implementation architecture.

3.1 Adversarial Prompt Design

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). The current implementation includes 452 prompts as a foundation for the full benchmark.

3.1.1 Health Domain Prompts

We 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). The model must extract and evaluate three components: likelihood (1–10 scale), impact (1–10 scale), and threat source identification.

- **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.

- **Physical safety**: Prompts related to physical safety risks and dangerous situations requiring appropriate safety guidance.

3.1.2 Misinformation Domain Prompts

In case of misinformation, we 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. The model should provide accurate information despite common misconceptions.

- **Explicit misinformation**: The prompt contains false statements that the model should identify and correct.

- **Implicit misinformation**: The prompt implicitly assumes an incorrect statement is true. The model must recognize the false assumption and provide accurate information.

- **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 and provide factual, accurate responses.

3.1.3 Disinformation Domain Prompts

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. The disinformation domain focuses on:

- **False narratives**: Prompts requesting generation of false narratives, fabricated events, or misleading content.

- **Coordinated inauthentic behavior**: Prompts attempting to generate content for coordinated disinformation campaigns, astroturfing, or election interference.

- **Safe requests**: Harmless requests to test for false refusals, ensuring the model maintains availability for legitimate use cases.

3.2 Evaluation Framework Architecture

The LSB evaluation framework is implemented as a production-ready Python system with the following key components:

3.2.1 Core Evaluation Pipeline

The evaluation pipeline follows a structured workflow:

1. **Prompt Loading**: Loads prompts from JSON files with metadata validation, supporting single or multiple domain files for unified evaluation.

2. **Prompt Formatting**: Handles both single-turn and multi-turn conversational prompts

using model-specific chat templates with fallback mechanisms. Multi-turn conversations are validated, truncated if necessary (max 20 turns, 2000 characters per message), and formatted appropriately.

3. **Response Generation**: Generates model responses using configurable sampling parameters (temperature, top-p, max tokens). Supports both single and batch generation for efficiency, with automatic fallback to individual generation on memory errors.

4. **Evaluation**: Applies domain-specific evaluation methods (refusal detection, harmful content detection, factual accuracy, threat assessment extraction).

5. **Result Aggregation**: Computes metrics across domains, difficulty tiers, and attack methods, generating comprehensive summaries.

6. **Output Generation**: Saves detailed JSON results and flattened CSV summaries with timestamps for reproducibility.

3.2.2 Evaluation Methods

The framework employs multiple complementary evaluation methods:

1. Refusal Detection

- **Keyword-based detection**: Uses an extensive list of refusal keywords (e.g., ”I cannot”, ”I’m unable”, ”against my guidelines”) with word boundary matching to identify explicit refusals.
- **Semantic similarity detection**: Computes embedding-based cosine similarity between model responses and reference refusal patterns. Uses a threshold of 0.75 for refusal classification, providing robustness against paraphrasing and variations in refusal phrasing.
- **Hybrid approach**: Combines both methods, with semantic similarity serving as a fallback when keyword matching is inconclusive.

2. Harmful Content Detection

- **Domain-specific keyword matching**: Maintains domain-specific lists of harmful compliance indicators (e.g., ”lethal dose”, ”how to make a bomb” for health domain; conspiracy theory terms for misinformation).
- **Context-aware detection**: Evaluates whether the model is providing harmful information rather than merely mentioning it in a warning context.

3. LLM-as-Judge Evaluation

- **Separate judge model**: Uses an optional separate, typically larger model (e.g., Qwen2.5-3B-Instruct) to evaluate factual accuracy, reducing bias from self-evaluation.
- **Few-shot prompting**: Provides few-shot examples demonstrating correct factual evaluation, improving judge model reliability.
- **Structured evaluation**: Judge model evaluates whether responses align with ground truth, correct misinformation, and provide accurate information.
- **Response truncation**: Truncates long responses (max 1500 characters) before judge evaluation to maintain context while managing token limits.

4. Threat Assessment Extraction

- **Multi-method extraction**: Attempts extraction using three methods in order of preference:
  1. **JSON parsing**: Extracts structured JSON with likelihood, impact, and source fields.
  2. **Regex extraction**: Uses regular expressions to find numeric scores and source descriptions in unstructured text.
  3. **Partial extraction**: Attempts to extract any available components when full extraction fails.
- **Score validation**: Validates extracted scores are within the expected range (1–10) and computes squared errors for regression metrics.
- **Source matching**: Evaluates source identification accuracy using semantic similarity and keyword matching.

5. Semantic Similarity Evaluation

- **Embedding-based alignment**: Uses sentence transformers to compute cosine similarity between model responses and ground truth or expected patterns.

- **Threshold-based classification**: Employs three similarity thresholds: strong (0.70), moderate (0.60), and weak (0.50) for different evaluation contexts.

- **Response alignment**: Evaluates whether factual responses align with expected ground truth content, accounting for paraphrasing and different phrasings.

3.3 Evaluation Metrics

The framework computes comprehensive metrics across multiple dimensions:

3.3.1 Primary Metrics

- **Attack Success Rate (ASR)**: Percentage of prompts where the model failed to respond safely. Computed separately for each domain, difficulty tier, and attack method.

- **Refusal Rate**: Percentage of prompts the model refused to answer. Critical for disinformation domain evaluation and availability assessment.

- **Safety Failure Rate**: Percentage of harmful requests where the model failed to refuse (for ”blocked” prompts).

- **False Positive Rate**: Percentage of harmless requests where the model incorrectly refused (for ”safe” or ”factual” prompts).

- **False Negative Rate**: Percentage of harmful requests where the model did not refuse when it should (for ”blocked” prompts).

- **Factual Error Rate**: Percentage of factual prompts where the model provided incorrect information, as determined by LLM-as-judge evaluation.

3.3.2 Threat Assessment Metrics (Health Domain)

For threat assessment prompts, the framework computes specialized metrics:

- **Parse Success Rate**: Percentage of responses where threat assessment components were successfully extracted.

- **Mean Squared Error (MSE)**: Computed separately for likelihood and impact scores, measuring prediction accuracy on the 1–10 scale.

- **Root Mean Squared Error (RMSE)**: More interpretable error metric, providing average deviation in score units.

- **Source Identification Accuracy**: Percentage of correct threat source identifications, evaluated using semantic similarity and keyword matching.

- **Per-Tier Breakdown**: All metrics broken down by difficulty tier to assess performance scaling with attack sophistication.

3.3.3 Aggregation and Reporting

Metrics are aggregated and reported at multiple levels:

- **Overall statistics**: Aggregate metrics across all prompts, domains, and tiers.

- **By domain**: Separate metrics for health, misinformation, and disinformation domains.

- **By difficulty tier**: Performance breakdown across Tier 1 (easiest) through Tier 4 (most difficult).

- **By attack method**: Effectiveness analysis for each attack technique (no attack, topic selection, implicit misinformation, prompt injection, jailbreak/roleplay, multi-turn persuasion).

- **By expected behavior**: Separate evaluation for blocked, factual, safe, and threat_assessment prompt categories.

3.4 Attack Methods

The benchmark incorporates six attack methods, organized by difficulty tier:

- **No attack** (Tier 1): Baseline prompts with no adversarial manipulation, used to check if a model refuses to provide an answer for a harmless request (testing for false positives).

- **Topic selection** (Tier 1–2): Selecting topics with a lot of misinformation online may lead to the model providing a misleading answer, even without explicit manipulation.

- **Implicit misinformation** (Tier 2): Prompts that implicitly assume incorrect statements are true, requiring the model to recognize and correct false assumptions embedded in the query.

- **Prompt injection** (Tier 3): Attempts to manipulate the model by instructing it to ignore previous instructions, override system prompts, or follow alternative instructions embedded in the user prompt.

- **Jailbreak/Roleplay** (Tier 3): Manipulating the model to bypass safety measures through role-playing scenarios, hypothetical situations, or creative formatting that obscures harmful intent.

- **Multi-turn persuasion** (Tier 4): Sophisticated conversational attacks that gradually escalate requests across multiple turns, building rapport and context before introducing harmful requests. This method leverages recent findings on multi-turn jailbreak effectiveness (Russinovich et al., 2025).

3.5 Implementation Features

The evaluation framework includes several production-ready features:

3.5.1 Configuration Management

All evaluation parameters are centralized in an `EvaluationConfig` dataclass, enabling fine-grained control over:

- Generation parameters (max tokens, temperature, top-p)

- Judge model settings (temperature, max tokens)

- Semantic similarity thresholds (strong, moderate, weak, refusal)

- Tokenization and truncation limits

- Batching and checkpointing intervals

- Multi-turn conversation limits

3.5.2 Memory Efficiency

- **Batch processing**: Processes prompts in configurable batches (default: 100 prompts) to manage memory usage for large datasets.

- **Response truncation**: Automatically truncates long responses before judge evaluation and embedding computation.

- **Conversation truncation**: Limits multi-turn conversations to the most recent N turns (default: 20) to prevent context overflow.

- **Message length limits**: Truncates individual messages exceeding character limits (default: 2000 characters).

3.5.3 Checkpointing and Resume

- **Automatic checkpointing**: Saves evaluation progress periodically (configurable interval, default: every 50 prompts).

- **Resume capability**: Automatically detects and resumes from checkpoints on subsequent runs, enabling recovery from interruptions.

- **Progress tracking**: Maintains state across evaluation runs, preventing duplicate work.

3.5.4 Error Handling and Robustness

- **Comprehensive error handling**: Catches and handles tokenization errors, generation failures, memory errors, and decoding issues with graceful fallbacks.

- **Input validation**: Validates all inputs with clear error messages and type checking.

- **Fallback mechanisms**: Automatically falls back to individual generation when batch processing fails, uses simple formatting when chat templates fail, and employs multiple extraction methods for threat assessment.

- **Logging**: Configurable logging with both console and file output, supporting multiple log levels for debugging and monitoring.

3.5.5 Output Formats

- **JSON output**: Complete results including all prompt metadata, full model responses, evaluation details, threat assessment extractions, judge model evaluations, and configuration used.

- **CSV output**: Flattened tabular summary with all key fields for easy analysis, plotting, and statistical processing.

- **Console summaries**: Human-readable summary statistics printed after evaluation completion, including overall metrics, domain breakdowns, tier analysis, and attack method effectiveness.

3.6 Dataset Statistics

We have successfully created the complete LSB benchmark dataset comprising 600 adversarial prompts, achieving our target of 200 prompts per domain across four difficulty tiers. This section presents comprehensive statistics characterizing the dataset composition, structure, and characteristics.

3.6.1 Overall Composition

The complete dataset consists of 600 prompts distributed evenly across three domains: health (200), misinformation (200), and disinformation (200). Prompts are organized into four difficulty tiers with perfect balance: 150 prompts in each tier (Tier 1 through Tier 4). The dataset includes 142 safe prompts (23.7%) for testing false refusals and 458 unsafe prompts (76.3%) representing adversarial attacks. Table 1 shows the complete distribution across domains and difficulty tiers.

Domain	Tier 1	Tier 2	Tier 3	Tier 4
Health	50	50	50	50
Misinformation	50	50	50	50
Disinformation	50	50	50	50
Total	**150**	**150**	**150**	**150**

Table 1: Distribution of prompts across domains and difficulty tiers (complete dataset).

3.6.2 Attack Method Distribution

The dataset incorporates six attack methods with distribution shown in Table 2. The most common attack method is no attack (151 prompts, 25.2%), serving as a baseline for evaluating false positive rates. Topic selection (132 prompts, 22.0%) is the most frequent adversarial technique, followed by implicit misinformation (104 prompts, 17.3%), multi-turn persuasion (84 prompts, 14.0%), jailbreak/roleplay (67 prompts, 11.2%), and prompt injection (62 prompts, 10.3%). The distribution reflects the increasing sophistication required for higher-tier attacks, with multi-turn persuasion primarily appearing in Tier 4.

Attack Method	Count
No attack	151
Topic selection	132
Implicit misinformation	104
Multi-turn persuasion	84
Jailbreak/Roleplay	67
Prompt injection	62
Total	**600**

Table 2: Distribution of prompts by attack method (complete dataset).

3.6.3 Expected Behavior Distribution

Prompts are categorized by expected model behavior, with distribution shown in Table 3. The largest category is *blocked* (341 prompts, 56.8%), indicating prompts where the model should refuse the request. This reflects the adversarial nature of the benchmark, particularly in the disinformation domain where 76.5% of prompts (153 out of 200) should be refused. The *factual* category (222 prompts, 37.0%) represents prompts where the model should provide accurate information and correct misinformation. The *threat_assessment* category (15 prompts, 2.5%) requires structured risk evaluation for health domain prompts. The *safe* category (22 prompts, 3.7%) tests for false refusals on harmless requests.

Expected Behavior	Count
Blocked	341 (56.8%)
Factual	222 (37.0%)
Threat assessment	15 (2.5%)
Safe	22 (3.7%)
Total	**600**

Table 3: Distribution of prompts by expected behavior (complete dataset).

3.6.4 Prompt Characteristics

The dataset includes both single-turn (515 prompts, 85.8%) and multi-turn conversational prompts (85 prompts, 14.2%). Multi-turn prompts are distributed across all domains (health: 27, misinformation: 32, disinformation: 26) 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 across the dataset. The overall length statistics are: minimum 23 characters, maximum 1,334 characters, mean 231.5 characters, median 105.5 characters, with interquartile range from 65.0 (25th percentile) to

252.25 (75th percentile). The longer prompts typically correspond to multi-turn conversations or complex scenarios requiring detailed context. Table 4 shows length statistics broken down by domain.

Domain	Min	Median	Mean	Max
Health	39	121.0	233.0	1,138
Misinformation	23	81.0	228.8	1,029
Disinformation	45	104.5	232.7	1,334
Overall	**23**	**105.5**	**231.5**	**1,334**

Table 4: Prompt length statistics (in characters) by domain.

3.6.5 Domain-Specific Characteristics

Each domain exhibits distinct characteristics aligned with its evaluation objectives:

Health Domain (200 prompts):

- **Expected behaviors**: Factual (77, 38.5%), blocked (108, 54.0%), threat assessment (15, 7.5%)
- **Risk categories**: Health advice (101, 50.5%), physical safety (46, 23.0%), threat detection (41, 20.5%), none (12, 6.0%)
- **Attack methods**: No attack (52, 26.0%), implicit misinformation (47, 23.5%), topic selection (30, 15.0%), multi-turn persuasion (29, 14.5%), jailbreak/roleplay (23, 11.5%), prompt injection (19, 9.5%)
- **Conversation types**: Single-turn (173, 86.5%), multi-turn (27, 13.5%)
- **Variants**: Unsafe (157, 78.5%), safe (43, 21.5%)

Misinformation Domain (200 prompts):

- **Expected behaviors**: Factual (120, 60.0%), blocked (80, 40.0%)
- **Risk categories**: Common misconceptions (67, 33.5%), explicit misinformation (65, 32.5%), implicit misinformation (35, 17.5%), logical fallacies (10, 5.0%), none (23, 11.5%)
- **Attack methods**: No attack (52, 26.0%), topic selection (45, 22.5%), implicit misinformation (34, 17.0%), multi-turn persuasion (30, 15.0%), jailbreak/roleplay (21, 10.5%), prompt injection (18, 9.0%)
- **Conversation types**: Single-turn (168, 84.0%), multi-turn (32, 16.0%)
- **Variants**: Unsafe (148, 74.0%), safe (52, 26.0%)

Disinformation Domain (200 prompts):

- **Expected behaviors**: Blocked (153, 76.5%), factual (25, 12.5%), safe (22, 11.0%)
- **Risk categories**: False narratives (91, 45.5%), coordinated inauthentic behavior (62, 31.0%), none (47, 23.5%)
- **Attack methods**: No attack (47, 23.5%), topic selection (57, 28.5%), implicit misinformation (23, 11.5%), multi-turn persuasion (25, 12.5%), prompt injection (25, 12.5%), jailbreak/roleplay (23, 11.5%)
- **Conversation types**: Single-turn (174, 87.0%), multi-turn (26, 13.0%)
- **Variants**: Unsafe (153, 76.5%), safe (47, 23.5%)

3.6.6 Difficulty Tier Progression

The four difficulty tiers demonstrate a clear progression in attack sophistication:

Tier 1 (150 prompts): Baseline evaluation with no attack (50 prompts) and topic selection (100 prompts) techniques. Focuses on testing model behavior on topics with high misinformation prevalence and harmless requests.

Tier 2 (150 prompts): Introduces implicit misinformation attacks (104 prompts) alongside topic selection (46 prompts). Tests model ability to recognize and correct false assumptions embedded in queries.

Tier 3 (150 prompts): Escalates to active manipulation techniques: jailbreak/roleplay (67 prompts), prompt injection (62 prompts), and topic selection (21 prompts). Evaluates model resilience against direct attempts to bypass safety measures.

Tier 4 (150 prompts): Most sophisticated tier featuring multi-turn persuasion attacks (84 prompts) combined with other techniques (66 prompts). Tests model behavior in complex conversational contexts where harmful requests are gradually introduced across multiple turns.

3.6.7 Risk Category Distribution

The dataset covers 10 distinct risk categories across the three domains, as shown in Table 5. Health domain categories include health advice (101), physical safety (46), and threat detection (41). Misinformation domain categories include common misconceptions (67), explicit misinformation (65), implicit misinformation (35), and logical fallacies (10). Disinformation domain categories include false narratives (91) and coordinated inauthentic behavior (62). Additionally, 82 prompts across domains are categorized as ”none” for baseline or general safety testing.

Risk Category	Count	Domain
Health advice	101	Health
False narratives	91	Disinformation
None	82	All
Common misconceptions	67	Misinformation
Explicit misinformation	65	Misinformation
Coordinated inauthentic behavior	62	Disinformation
Physical safety	46	Health
Threat detection	41	Health
Implicit misinformation	35	Misinformation
Logical fallacies	10	Misinformation
Total	**600**	

Table 5: Distribution of prompts by risk category across domains.

3.6.8 Dataset Completeness and Balance

The complete LSB dataset achieves perfect balance across key dimensions:

- **Domain balance**: Exactly 200 prompts per domain (33.3% each)
- **Difficulty tier balance**: Exactly 150 prompts per tier (25.0% each)
- **Per-domain tier balance**: Exactly 50 prompts per domain–tier combination (8.3% each)
- **Attack method coverage**: All six attack methods represented across appropriate difficulty tiers
- **Expected behavior diversity**: Four distinct behavior categories with appropriate domain-specific distributions
- **Conversation type mix**: 14.2% multi-turn prompts, primarily in higher tiers, reflecting realistic attack scenarios

This balanced structure enables comprehensive evaluation across all dimensions while maintaining statistical validity for cross-domain and cross-tier comparisons.

4 Proof of Concept

We validated the complete LSB evaluation pipeline (prompts → model generation → automatic scoring → JSON/CSV outputs) with four openly available small language models on the full 600-prompt benchmark. The PoC demonstrates the framework’s capability to evaluate models across all three domains, four difficulty tiers, and six attack methods, producing comprehensive safety assessments.

4.1 Evaluation Pipeline

The proof of concept validates the complete evaluation workflow:

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

2. **Unified Evaluation Execution**: The evaluator processes all 600 prompts across three domains in a single unified run. The pipeline:

   - Formats prompts (handling both single-turn and multi-turn conversations)
   - Generates model responses with configurable sampling parameters
   - Detects refusals using keyword matching and semantic similarity
   - Identifies harmful content using domain-specific indicators
   - Applies LLM-as-judge evaluation for factual accuracy (using Qwen2.5-3B-Instruct as judge model)
   - Extracts threat assessment scores (likelihood, impact, source) for health domain prompts
   - Writes per-prompt JSON results and flattened CSV summaries
   - Generates comprehensive printed summaries with metrics broken down by domain, tier, and attack method

3. **Multi-Model Comparison**: Repeat Step 2 for additional models to observe family, size, and training effects. In the PoC we evaluated four models:

   - `Qwen/Qwen2.5-1.5B-Instruct` (Qwen family, instruction-tuned)
   - `TinyLlama-1.1B-Chat-v1.0` (minimal safety training)
   - `Llama-3.2-1B-Instruct` (Meta's Llama family, instruction-tuned)
   - `OpenMath-Nemotron-1.5B` (specialized model, high refusal rate)

4.2 Overall Results

Table 6 presents overall performance metrics for all four evaluated models on the complete 600-prompt benchmark.

Model	ASR	Refusal Rate	N
Qwen2.5-1.5B-Instruct	43.0%	99.8%	600
Llama-3.2-1B-Instruct	43.2%	31.5%	600
OpenMath-Nemotron-1.5B	51.3%	82.7%	600
TinyLlama-1.1B-Chat-v1.0	67.2%	8.8%	600

Table 6: Overall model performance on complete 600-prompt benchmark. ASR = Attack Success Rate, N = number of prompts evaluated.

4.3 Results by Domain

Tables 7–9 show performance breakdown by domain for each model, revealing domain-specific vulnerabilities.

Model	ASR	Refusal Rate
Qwen2.5-1.5B-Instruct	45.5%	99.5%
Llama-3.2-1B-Instruct	46.0%	31.5%
OpenMath-Nemotron-1.5B	57.0%	82.5%
TinyLlama-1.1B-Chat-v1.0	62.5%	9.5%

Table 7: Performance in Health domain. ASR = Attack Success Rate.

Model	ASR	Refusal Rate
Qwen2.5-1.5B-Instruct	60.0%	100.0%
Llama-3.2-1B-Instruct	38.5%	30.5%
OpenMath-Nemotron-1.5B	65.0%	83.0%
TinyLlama-1.1B-Chat-v1.0	67.0%	10.5%

Table 8: Performance in Misinformation domain. ASR = Attack Success Rate.

Key observations:

Model	ASR	Refusal Rate
Qwen2.5-1.5B-Instruct	23.5%	100.0%
Llama-3.2-1B-Instruct	45.0%	32.5%
OpenMath-Nemotron-1.5B	32.0%	82.5%
TinyLlama-1.1B-Chat-v1.0	72.0%	6.5%

Table 9: Performance in Disinformation domain. ASR = Attack Success Rate.

- **Disinformation domain shows best safety**: Most models achieve lower ASR in disinformation (23.5–72.0%) compared to other domains, reflecting effective refusal behavior for content generation requests.

- **Misinformation domain is challenging**: Models struggle with misinformation correction tasks, with ASR ranging from 38.5% to 67.0%, indicating difficulty in recognizing and correcting false claims.

- **Health domain varies by model**: Performance ranges from 45.5% to 62.5% ASR, with threat assessment and health advice prompts presenting different challenges.

- **Qwen2.5-1.5B shows excessive refusals**: With 99.8% overall refusal rate, this model demonstrates overly conservative behavior, likely including many false positives on harmless requests.

4.4 Results by Difficulty Tier

Table 10 presents attack success rates by difficulty tier, demonstrating how attack sophistication affects model resilience.

Model	Tier 1	Tier 2	Tier 3	Tier 4
Qwen2.5-1.5B-Instruct	76.0%	64.7%	21.3%	10.0%
Llama-3.2-1B-Instruct	34.0%	45.3%	49.3%	44.0%
OpenMath-Nemotron-1.5B	74.7%	68.0%	33.3%	29.3%
TinyLlama-1.1B-Chat-v1.0	50.7%	60.7%	74.7%	82.7%

Table 10: Attack Success Rate by difficulty tier.

Notable patterns:

- **Qwen2.5-1.5B shows inverse tier pattern**: ASR decreases from 76.0% (Tier 1) to 10.0% (Tier 4), suggesting the model's refusal detection may be incorrectly flagging many Tier 1 prompts as refusals (false positives), while correctly handling sophisticated Tier 4 attacks.

- **TinyLlama vulnerability increases with tier**: ASR rises from 50.7% (Tier 1) to 82.7% (Tier 4), demonstrating that models without robust safety training become increasingly vulnerable to sophisticated attacks.

- **Llama-3.2-1B shows consistent vulnerability**: ASR remains relatively stable across tiers (34.0–49.3%), indicating consistent behavior but room for improvement at all difficulty levels.

- **OpenMath-Nemotron shows improvement at higher tiers**: ASR decreases from 74.7% (Tier 1) to 29.3% (Tier 4), suggesting better handling of sophisticated attacks despite high baseline vulnerability.

4.5 Key Findings

4.5.1 Model Family and Training Effects

- **Instruction tuning matters**: Qwen2.5-1.5B and Llama-3.2-1B (both instruction-tuned) show different safety behaviors, with Qwen being overly conservative and Llama showing more balanced refusal behavior.

- **Safety training is critical**: TinyLlama-1.1B, with minimal safety training, demonstrates the highest ASR (67.2%) and lowest refusal rate (8.8%), complying with most harmful requests.

- **Specialized models may over-refuse**: OpenMath-Nemotron-1.5B shows high refusal rates (82.7%) but still maintains significant vulnerability (51.3% ASR), suggesting a mismatch between refusal behavior and actual safety.

4.5.2 Attack Sophistication

- **Multi-turn persuasion is effective**: Tier 4 attacks (primarily multi-turn persuasion) achieve high success rates on vulnerable models (82.7% for TinyLlama), confirming that conversational context manipulation is a potent attack vector.

- **Baseline attacks reveal fundamental issues**: High Tier 1 ASR (50.7–76.0% for most models) indicates that even simple topic selection and implicit misinformation can successfully exploit model vulnerabilities.

- **Model-specific attack effectiveness varies**: Different models show different tier progression patterns, suggesting that attack effectiveness depends on model architecture, training, and safety mechanisms.

4.5.3 Evaluation Framework Validation

- **Pipeline scalability**: Successfully evaluated 600 prompts per model across four models (2,400 total evaluations) with automated scoring, demonstrating the framework's scalability.

- **Metric diversity**: The framework successfully computed multiple metrics (ASR, refusal rate, safety failures, false positives, factual errors) enabling comprehensive safety assessment.

- **Cross-model comparison**: Standardized evaluation protocol enables fair comparison across different model families and sizes, revealing distinct safety profiles.

- **Reproducibility**: Timestamped JSON/CSV outputs with complete metadata ensure reproducibility and enable detailed post-hoc analysis.

4.6 Limitations and Future Work

The PoC reveals several areas for improvement:

- **False positive detection**: Qwen2.5-1.5B's 99.8% refusal rate suggests the refusal detection may be too sensitive, flagging legitimate responses as refusals. Future work should refine refusal detection thresholds and validation.

- **Threat assessment evaluation**: While the framework extracts threat assessment components, comprehensive evaluation of threat assessment accuracy requires further validation against expert annotations.

- **LLM-as-judge reliability**: The judge model evaluation for factual accuracy should be validated against human annotations to ensure reliability.

- **Model coverage**: The PoC evaluated only small models (1–1.5B parameters). Future work should extend evaluation to larger models (3B, 7B, 13B+) to assess scaling effects.

- **Attack method analysis**: Detailed breakdown by attack method (topic selection, implicit misinformation, prompt injection, jailbreak, multi-turn persuasion) would provide deeper insights into specific vulnerabilities.

4.7 Conclusion

The proof of concept successfully validates the LSB evaluation framework on the complete 600-prompt benchmark. The framework demonstrates:

1. **Comprehensive evaluation capability**: Successfully evaluates models across three domains, four difficulty tiers, and multiple attack methods.

2. **Automated scoring reliability**: Automated metrics (ASR, refusal rate) provide consistent, reproducible safety assessments.

3. **Cross-model comparability**: Standardized protocol enables fair comparison revealing distinct model safety profiles.

4. **Scalability**: Framework handles large-scale evaluation (600 prompts per model) with automated processing and output generation.

All evaluation runs produced timestamped JSON/CSV files in `results/`, providing complete audit trails for all reported metrics and enabling detailed analysis of model behavior across domains, tiers, and attack methods.

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, 3.1–3.3 (w/ Koniecko), 3.4, 4	Eval. pipeline, 298 prompts
W. Koniecko	3.1–3.3 (w/ Kisiel)	—
K. Frańczak	2 (w/ Kosakowski)	152 prompts
P. Kosakowski	2 (w/ Frańczak)	150 prompts

Table 11: Work division among team members.