Position Paper: Just as Humans Need Vaccines, So Do Models - Immunization to Combat Falsehoods

Shaina Raza*
Vector Institute
Toronto, Canada

Rizwan Qureshi
University of Central Florida
Orlando, USA

Azib Farooq
University of Cincinnati
Cincinnati, USA

Marcelo Lotif
Vector Institute
Toronto, Canada

Aman Chadha†
Independent Researcher
USA

Deval Pandya
Vector Institute
Toronto, Canada

Christos Emmanouilidis
University of Groningen
The Netherlands

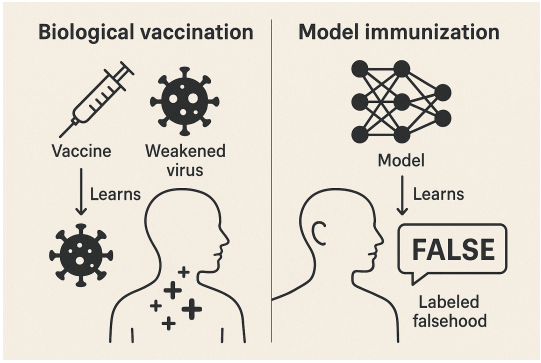
***Abstract*—Large language models (LLMs) reproduce misinformation by learning the linguistic patterns that make falsehoods persuasive, ssuch as hedging, false presuppositions, and citation fabrication, rather than merely memorizing false facts. We propose model immunization: supervised fine-tuning on curated (false claim, correction) pairs injected as small "vaccine doses" (5–10% of tokens) alongside truthful data. Unlike post-hoc filtering or preference-based alignment, immunization provides direct negative supervision on labeled falsehoods. Across four open-weight model families, immunization improves TruthfulQA accuracy by 12 points and misinformation rejection by 30 points with negligible capability loss. We outline design requirements, which includes, dosage, labeling, quarantine, diversity and call for standardized vaccine corpora and benchmarks that test generalization, making immunization a routine component of responsible LLM development. The project webpage is available at project page and the code repository at GitHub.**

***Index Terms*—large language models, misinformation, bias, generative AI, responsible AI**

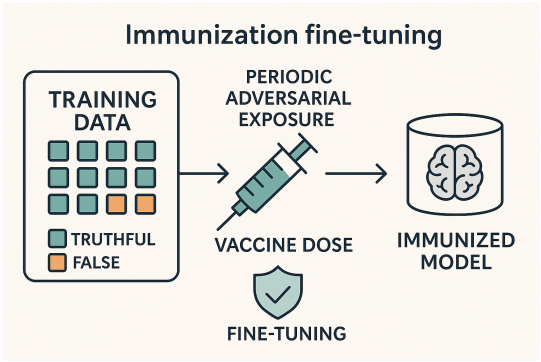
I. Introduction

When a language model asserts that "vaccines cause autism" [1] or that "the 2020 U.S. election was stolen" [2], it is not simply retrieving a memorized false fact. It has learned something deeper: the linguistic conventions through which such claims are constructed and made persuasive. These include hedging patterns ("some people say..."), presuppositional triggers that smuggle in false premises, citation fabrications that mimic authoritative sourcing, and rhetorical structures that lend credibility to baseless claims. The problem, in other words, is not merely one of factual knowledge but of linguistic competence, which means the models have learned *how* misinformation sounds, and they reproduce those patterns regardless of truth value.

Current approaches to improving model truthfulness do not directly address this phenomenon. Post-hoc detection systems attempt to catch misinformation after generation [3]; for example, they treat symptoms rather than causes. Reinforcement learning from human feedback (RLHF) provides indirect pressure toward truthfulness via reward signals [4],



(a) Biological vaccination vs. model immunization: controlled exposure builds resistance.



(b) Immunization fine-tuning: periodic injection of labeled falsehoods (5–10%) alongside truthful data.

Fig. 1: The immunization analogy. Just as biological vaccines use controlled exposure to weakened pathogens to train immune responses, model immunization uses controlled exposure to labeled falsehoods to train rejection responses.

but does not give models explicit supervision on what makes a claim false; therefore, truthfulness is diluted across helpfulness, harmlessness, and stylistic preferences. Adversarial training harden models against input perturbations [5], such as typos, paraphrases, adversarial suffixes, but does not address the semantic content of misinformation. A model robust to character-level attacks may still enthusiastically endorse false

*Corresponding author. Email: shaina.raza@torontomu.ca
†Work done outside of Amazon.

claims presented in clean text.

We propose a more direct intervention: **supervised immunization against misinformation**. The core idea is simple but represents a conceptual shift grounded in psychological inoculation theory [6], which demonstrates that controlled exposure to weakened forms of persuasive attacks can build cognitive resistance in humans [7]. Rather than filtering false data out of training corpora, we deliberately include it, but with explicit negative labels and paired corrections. Just as **biological vaccines** use controlled exposure to weakened pathogens to train immune responses (Figure 1a), **model immunization** uses controlled exposure to labeled falsehoods to train rejection responses. In practice, this means periodically injecting small "doses" of fact-checked false claims (5–10% of fine-tuning tokens) alongside truthful data (Figure 1b), with each falsehood explicitly marked as false and paired with correct information. This reframes the role of false data in alignment. Standard practice treats misinformation as contamination; we argue it should be treated as *evasive signal*, information if properly labeled and governed, teaches models what to avoid. The approach is also analogous to how toxic language datasets are used to train content filters [8]: we do not hide offensive content from the model but show it explicitly, with labels, so the model learns to recognize and reject it.

Our **position** rests on three claims. First, *truthfulness requires negative supervision*. Preference-based methods like RLHF encode factuality as one objective among many; models need explicit examples of falsehoods labeled as such: direct negative signal, not indirect preference pressure. Second, *misinformation is fundamentally a linguistic phenomenon*. False claims exhibit characteristic patterns that models learn and reproduce; immunization should target these patterns, not just specific false propositions. Third, *the AI and NLP communities need new infrastructure* to make immunization practical: standardized vaccine corpora, benchmarks testing robustness to unseen misinformation, and evaluation protocols measuring generalization across languages and domains.

II. The Linguistic Case for Misinformation Research

Misinformation is fundamentally a *linguistic* phenomenon. Beyond incorrect facts, false claims exhibit recurring discourse patterns that make them persuasive, including hedged assertions, false presuppositions, citation fabrication, emotive amplification, and false balance. These patterns recur across domains such as health, politics, and science, and are learned by language models from web-scale corpora. Table I summarizes six such linguistic signatures commonly observed in misinformation. Crucially, these patterns generalize across claims: removing specific false statements from training data does not remove the underlying discourse templates. As a result, models may correctly answer factual queries while still reproducing misinformation when prompted with persuasive linguistic framing. This observation motivates training interventions that target *patterns*, not just propositions. Improving truthfulness therefore requires teaching models to recognize and reject misinformation-associated linguistic features, rather than relying solely on fact recall or post-hoc filtering.

TABLE I: Linguistic patterns characteristic of misinformation. Each represents a learnable feature that models acquire from training data.

Pattern	Description and Example
Hedged assertion	Weasel words avoiding commitment: "Some experts believe...", "It has been reported that..."
False presupposition	Smuggles false premise into discourse: "When did the government admit the cover-up?"
Citation fabrication	Mimics scholarly sourcing: "According to a Stanford study..." (nonexistent)
Emotive amplification	Substitutes affect for evidence: "The SHOCKING truth they don't want you to know"
False balance	Presents fringe views as equally credible: "While most scientists say X, others argue Y"
Temporal manipulation	Implies causation through proximity: "Shortly after the vaccine rollout, deaths increased"

III. Immunization as Negative Supervision

Having established that misinformation is a linguistic phenomenon, we now present model immunization as a direct intervention targeting the patterns described above.

A. The Core Proposal

Model immunization is supervised fine-tuning on a curated corpus of (false claim, correction) pairs, where each false claim carries an explicit negative label and the model learns to reject it or generate the correction rather than endorse the claim. The approach treats fact-checked falsehoods not as contamination to be removed but as training signal to be harnessed, which is analogous to how exposure to weakened pathogens trains biological immune systems. The framework requires **four** design choices.

(1) *Dosage* determines what fraction of fine-tuning tokens come from the vaccine corpus. Too little yields no measurable effect; too much risks teaching the model to generate misinformation rather than reject it. Preliminary experiments suggest 5–10% strikes an appropriate balance, though the optimal dose likely varies with model scale, domain, and the diversity of the vaccine corpus.

(2) *Labeling* ensures the model learns falsehoods as negative examples rather than internalizing them as knowledge. Each claim is paired with an explicit "this is false" marker and a correction providing the factual alternative. The training objective penalizes the model for reproducing or endorsing the false claim and rewards it for generating the correction or refusing to engage. Without clear labeling, the distinction between vaccine data and knowledge data collapses, and immunization becomes poison.

(3) *Quarantine* keeps vaccine data strictly separate from the model's knowledge base throughout the training pipeline. Falsehoods must never be retrievable as facts; they exist only as patterns to reject. This separation is both technical, distinct data pipelines, separate storage, clear provenance tracking,

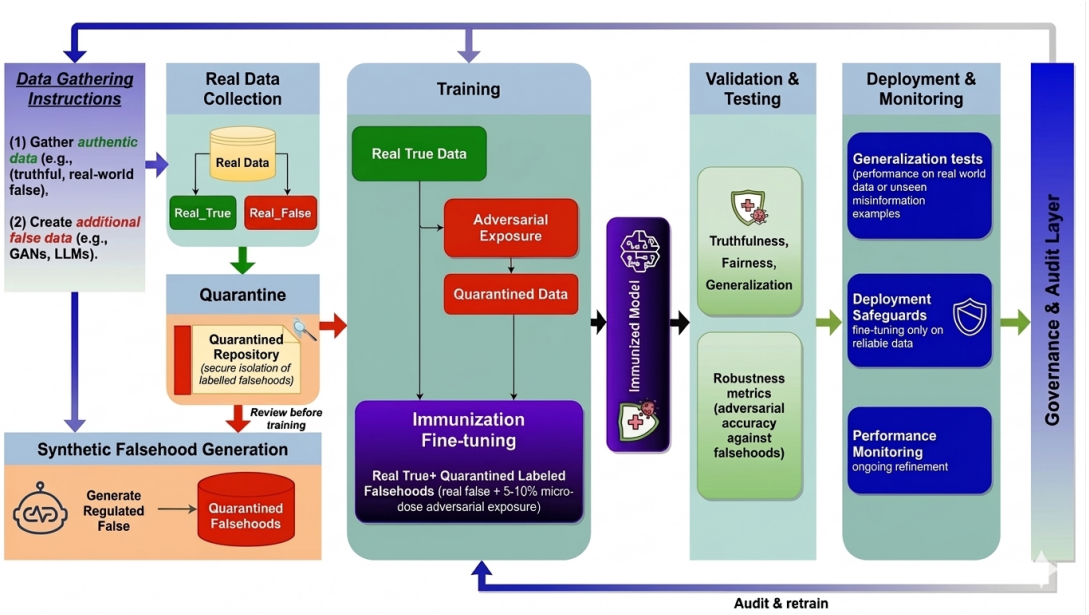


Fig. 2: The immunization pipeline. Data curation separates truthful and false examples; falsehoods enter a quarantined repository with fact-checker verification before controlled injection during fine-tuning. Validation tests truthfulness and robustness to held-out misinformation; deployment includes continuous monitoring with feedback for iterative refinement. A governance layer ensures accountability and auditability throughout.

and governance-related, with audit trails documenting every falsehood's source, verification, and use in training.

(4) *Diversity* requires coverage across misinformation types and linguistic patterns, not just specific claims. A vaccine corpus containing only health misinformation will not immunize against political falsehoods; one containing only English claims will not transfer to other languages. Effective immunization demands breadth across domains (health, politics, science, history), across languages, and across the linguistic signatures described in Section II. The goal is pattern-level immunity, not instance-level memorization.

Figure 2 illustrates the complete pipeline. The process begins with data curation, where authentic data and real-world falsehoods are collected from reliable sources, for example, fact-checking organizations, debunking databases, and known hoax repositories. These are augmented with synthetic falsehoods generated to fill coverage gaps, ensuring the vaccine corpus spans the full space of misinformation patterns. All false claims enter a quarantined repository where they are reviewed, labeled, and paired with corrections before any training use.

During immunization fine-tuning, the model receives a controlled mixture of standard training data and vaccine doses. The quarantined falsehoods are injected periodically, not concentrated in a single phase but distributed across training to reinforce rejection patterns throughout. Validation testing then evaluates both truthfulness (*does the model reject known falsehoods?*) and generalization *(does it reject novel falsehoods exhibiting similar patterns?*). Deployment includes ongoing monitoring for emerging misinformation and mechanisms for "**booster**" updates when new false narratives appear. Unlike standard supervised setups, immunization fine-tuning exposes the model only to **false claims with fact-checker corrections**, it is equivalent to training response behavior rather than true–false classification.

B. Distinguishing Immunization from Related Approaches

Model immunization occupies a distinct niche in the landscape of misinformation defenses. Figure 3 positions it relative to adversarial training, RLHF, and post-hoc detection along two dimensions: when in the model lifecycle the intervention occurs, and what type of signal it provides for factuality. The comparison with **RLHF** is particularly important, where human raters prefer truthful responses, but this signal is diluted across helpfulness, and other criteria, where truthfulness competes with multiple objectives [4]. Immunization provides *direct* supervision: each false claim is explicitly labeled and paired with a correction.

Adversarial training targets a different problem: robustness to input perturbations like typos and adversarial suffixes [5]. This hardens models against *how* inputs are presented, not *what* they claim. A model resistant to typo-based attacks may still endorse misinformation in clean prose. The two approaches are complementary. **Data decontamination**, such as filtering falsehoods from training corpora, removes specific claims but does not teach rejection as a skill [9]. A decontaminated model has never seen certain falsehoods; it has not learned to refuse them. Immunization inverts this logic: include falsehoods with explicit negative labels so the model learns what to reject. **Post-hoc detection** catches misinformation

(a) Comparison by input type, goal, and factuality signal.

Technique	Input	Goal	Signal
Adversarial training	perturbations	robustness	none
RLHF	preferences	alignment	indirect
Immunization	**falsehoods**	**truthfulness**	**direct**
Post-hoc detection	outputs	filtering	reactive

(b) Lifecycle timeline: when each defense applies.

Fig. 3: Misinformation defense techniques across LLM lifecycle. Immunization provides direct negative supervision on labeled falsehoods during fine-tuning, distinct from indirect preference signals (RLHF) and reactive output filtering (post-hoc detection).

after generation [10]. This is reactive, as errors must first be produced, then filtered. Immunization is preventive that reduces the burden on downstream classifiers by addressing the root cause.

C. Governance Requirements

Using false data as a training signal raises legitimate concerns that require structured governance. Without safeguards,"immunization" could become a vector for data poisoning, deliberately teaching models misinformation under the guise of inoculation. We address this through five governance principles. First, *transparency* requires documenting every use of false data in training, ensuring each quarantined falsehood is traceable from source to final model with audit logs recording provenance and labeling decisions [11]. Second, *no promotion of false content* ensures every curated false statement is treated as a negative training signal, with the training objective penalizing reproduction or endorsement of false claims while rewarding corrections or refusals [12]. Third, *alignment with human values* focuses curation on clearly discredited falsehoods with broad consensus such as dangerous health misinformation, debunked conspiracy theories, and verified hoaxes, while contested claims require additional human oversight [13]. Fourth, *preventing abuse* distinguishes responsible immunization from covert poisoning through open documentation, shared protocols, and use of publicly verifiable fact-checked datasets. Finally, *continuous accountability* establishes mechanisms for ongoing oversight, including public reporting channels that allow users to flag cases where immunization may have failed, feeding back into corpus refinement and model updates [14].

IV. A Research Agenda for the AI and NLP Communities

Model immunization cannot succeed without new infrastructure. Current resources are inadequate in three respects: benchmarks measure the wrong capabilities, appropriate training corpora do not exist at scale, and evaluation protocols ignore generalization. This section outlines what the community must build.

A. Benchmarks That Test Resistance, Not Recall

Existing factuality benchmarks test whether models *know* true facts, for example, TruthfulQA [15] measures avoidance of false claims, FActScore [16] measures factual precision. But knowing facts and resisting misinformation are different capabilities. A model might correctly answer "Is the earth round?" while generating flat-earth content when prompted to argue for it, or refuse to state "Vaccines cause autism" while hedging when asked about vaccine-autism relationships. We need benchmarks that test **defensive** capabilities: **endorsement resistance** (does the model refuse to validate false claims?), **elicitation resistance** (does it refuse prompts designed to generate misinformation?), **presupposition detection** (does it correct false premises rather than accepting them?), and **cross-type generalization** (does resistance transfer to unseen misinformation categories?). Such a benchmark would complement existing measures by testing defense, not just recall.

B. Vaccine Corpora That Enable Systematic Training

Ad-hoc falsehood datasets produce fragmentation, for example, variable quality, inconsistent labeling, irreproducible results. The community needs shared resources analogous to FEVER [17], but designed for negative supervision. Effective vaccine corpora require: **multilingual coverage** from independent fact-checkers across languages; **linguistic annotation** tagging patterns from Section II (hedging, presupposition, citation fabrication); **domain stratification** by topic (health, politics, science) for cross-domain studies; and **synthetic augmentation** via template-based generation to span possible misinformation beyond observed instances. Building such corpora requires collaboration between AI and NLP researchers, fact-checkers, and domain experts, but without standardized resources, immunization research will remain fragmented.

C. Evaluation Protocols That Demand Generalization

Demonstrating that a model rejects the specific falsehoods it was trained on proves little. Memorization is easy; generalization is hard [18]. If immunization only teaches models to reject the 500 claims in the vaccine corpus, it offers no protection against the boundless space of misinformation in the wild. Evaluation protocols must therefore require generalization tests as standard, not optional. **Held-out claim types** test whether immunization against one category (e.g., health misinformation) transfers to another (e.g., political misinformation) absent from training. Evaluation protocols must require generalization tests as standard, not optional. Without testing transfer across domains, linguistic framings, languages, and time, immunization risks overfitting to specific false claims rather than learning robust resistance patterns. **Temporal generalization** tests whether immunization against

established misinformation helps with newly emerging false claims. Without these generalization requirements, immunization research risks producing a literature of overfitted results that collapse upon deployment. The benchmarks and corpora described above must be designed with generalization testing built in, not as an afterthought.

D. Integration with Standard Development Pipelines

Finally, immunization should become a standard stage in responsible LLM development (not an optional enhancement but a default component of training pipelines). This raises practical questions that require empirical investigation.

Ordering effects: Integration with standard development pipelines raises practical questions, including the ordering of immunization relative to RLHF, risks of over-refusal, and the need for periodic booster updates as misinformation evolves. We leave systematic study of these trade-offs to future work. **Interaction with instruction-following**: Aggressive immunization might cause over-refusal, where models refuse legitimate queries that superficially resemble misinformation. The sensitivity-specificity trade-off requires characterization: how much immunization is too much? Does the answer depend on deployment context? **Maintenance and boosters**: New misinformation emerges continuously. How frequently must models receive "booster" updates with newly emerged false claims? Can incremental immunization be effective, or does each update require full fine-tuning? What infrastructure supports rapid response to emerging misinformation threats? These questions define a research program, not a solved problem. We offer immunization as a framework for investigation, not a finished solution.

V. Experimental Evaluation

We evaluate immunization through three experiments: efficacy across model families, cross-domain generalization, and dosage ablation.

A. Setup

Vaccine corpus. We compile 1,611 fact-checked false claims from LIAR [19] (politics), Health Feedback [20] (health), and Snopes/AFP [21], [22] (science). Each claim is paired with a fact-checker–derived correction. Vaccine examples are mixed with 30,609 truthful QA pairs from SQuAD [23] at 5% vaccine ratio.

Models. We test four model families: Phi-3-mini-4k-instruct (3.8B) [24], Llama-2-7B-Chat [25], Mistral-7B-Instruct-v0.2 [26], and Llama-3-8B-Instruct [27].

Evaluation. We measure TruthfulQA accuracy (817 questions) and rejection rate on 200 held-out false claims.

B. Results

Exp. 1: Efficacy. Table II shows consistent improvements across all models: +12.4pp on TruthfulQA and +30.2pp on misinformation rejection. MMLU remains stable (–0.5pp average), indicating no degradation of general capabilities.

TABLE II: Immunization efficacy across model families (5% vaccine dosage).

	TruthfulQA		Rejection	
Model	Base	Imm.	Base	Imm.
Phi-3-mini	38.2	47.6 +9.4	41.5	68.0 +26.5
Llama-2-7B	35.1	48.3 +13.2	43.0	74.5 +31.5
Mistral-7B	42.3	55.8 +13.5	47.5	79.0 +31.5
Llama-3-8B	44.7	58.2 +13.5	51.0	82.5 +31.5
Average	40.1	52.5 +12.4	45.8	76.0 +30.2

TABLE III: Cross-domain transfer (health-only training).

Test Domain	Rejection	Gap
Health (in-domain)	84.2%	—
Political (held-out)	61.8%	–22.4
Science (held-out)	68.5%	–15.7
Avg. held-out	65.2%	–19.0

Exp. 2: Generalization. To test cross-domain transfer, we train on health misinformation only and evaluate on held-out domains (Table III). The model rejects 65% of unseen-domain misinformation versus 84% in-domain—partial but meaningful transfer suggesting pattern-level learning beyond specific claims.

TABLE IV: Dosage ablation (Mistral-7B-Instruct).

Dosage	TruthfulQA	Rejection	MMLU
0% (base)	42.3	47.5	58.4
2%	47.8 +5.5	64.5 +17.0	58.2
5%	55.8 +13.5	79.0 +31.5	57.9
10%	57.2 +14.9	83.5 +36.0	57.1
20%	56.8 +14.5	85.0 +37.5	55.2

Exp. 3: Dosage. Table IV shows a clear dose-response relationship. Gains plateau after 10%, while 20% dosage degrades MMLU by 1.9pp. This supports our 5–10% recommendation: sufficient signal without capability loss.

VI. Limitations and Open Questions

We acknowledge significant gaps in the current proposal and resist the temptation to oversell. First,the argument is primarily **conceptual;** comprehensive empirical validation across model scales and domains remains future work. Preliminary experiments suggest immunized models show improved truthfulness without degrading general capabilities [28], but these results use small models and limited vaccine corpora. Second, the **scalability challenge** is fundamental. The space of possible falsehoods is unbounded, and new misinformation emerges faster than any corpus can track. Immunization cannot cover all possible false claims; it must rely on pattern generalization. Third, **over-refusal** poses a practical risk. Aggressive immunization might teach models to refuse legitimate queries that superficially resemble misinformation. A model trained extensively on health misinformation might refuse to discuss vaccines at all, even for legitimate educational purposes. Fourth,

governance introduces normative complexity. Determining what counts as misinformation is sometimes contested [29]. We focus on empirically debunked claims with broad scientific consensus, but edge cases abound. We do not resolve this tension but flag it as requiring ongoing deliberation. Lastly, **cross-lingual coverage is severely limited.** Fact-checking resources concentrate in high-resource languages; extending immunization to low-resource languages faces data scarcity challenges common across AI and NLP. Multilingual immunization is technically possible but practically constrained by the availability of verified falsehoods in diverse languages.

VII. Conclusion

We showed that misinformation in LLMs arises not only from incorrect facts, but from learned linguistic patterns that make falsehoods persuasive. We introduced model immunization, a supervised fine-tuning approach that provides direct negative supervision on labeled false claims paired with corrections. Across multiple open-weight models, immunization substantially improves truthfulness and misinformation rejection without degrading general capabilities. This re-frames false data from contamination to training signal when properly labeled and governed. We argue that negative supervision should be treated as a first-class tool in responsible LLM development and incorporated alongside existing alignment methods.

Acknowledgment

Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute. This research was funded in part by the European Union's Horizon Europe research and innovation programme under the AIXPERT project (Grant Agreement No. 101214389), which aims to develop an agentic, multi-layered, GenAI-powered framework for creating explainable, accountable, and transparent AI systems.

References

[1] C. Wardle and H. Derakhshan, "Information disorder: Toward an interdisciplinary framework for research and policymaking," *Council of Europe report*, vol. 27, pp. 1–107, 2017.

[2] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: Analyzing language in fake news and political fact-checking," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2931–2937.

[3] X. Zhou, A. Sharma, A. X. Zhang, and T. Althoff, "Correcting misinformation on social media with a large language model," *arXiv preprint arXiv:2403.11169*, 2024.

[4] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[5] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[6] W. J. McGuire and D. Papageorgis, "Resistance to persuasion conferred by active and passive prior refutation of the same and alternative counterarguments," *Journal of Abnormal and Social Psychology*, vol. 63, no. 2, pp. 326–332, 1961.

[7] J. Roozenbeek, S. van der Linden *et al.*, "Psychological inoculation improves resilience against misinformation on social media," *Science Advances*, vol. 8, no. 34, 2022.

[8] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "Realtoxicityprompts: Evaluating neural toxic degeneration in language models," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 3356–3369. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.301/

[9] J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner, "Documenting large webtext corpora: A case study on the Colossal Clean Crawled Corpus," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 1286–1305.

[10] Z. Guo, M. Schlichtkrull, and A. Vlachos, "A survey on automated fact-checking," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 178–206, 2022.

[11] Y. Chen *et al.*, "Towards transparent ai: A survey on explainable large language models," 2025.

[12] S. Raza and C. Ding, "Fake news detection based on news content and social contexts: a transformer-based approach," *International Journal of Data Science and Analytics*, vol. 13, no. 4, pp. 335–362, 2022.

[13] B. Larsen and V. Dignum. (2024, Oct.) Ai value alignment: How we can align artificial intelligence with human values. World Economic Forum. Accessed 2025-07-23.

[14] M. Busuioc, "Accountable artificial intelligence: Holding algorithms to account," *Public administration review*, vol. 81, no. 5, pp. 825–836, 2021.

[15] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," *arXiv preprint arXiv:2109.07958*, 2022.

[16] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "FActScore: Fine-grained atomic evaluation of factual precision in long form text generation," *arXiv preprint arXiv:2305.14251*, 2023.

[17] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "Fever: A large-scale dataset for fact extraction and verification," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 809–819, dataset available at: https://fever.ai/dataset/fever.html.

[18] R. Volpi, H. Namkoong, O. Sener, J. C. Duchi, V. Murino, and S. Savarese, "Generalizing to unseen domains via adversarial data augmentation," *Advances in neural information processing systems*, vol. 31, 2018.

[19] W. Y. Wang, ""liar, liar pants on fire": A new benchmark dataset for fake news detection," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 2017, pp. 422–426.

[20] Health Feedback, "Health feedback: A global network of scientists sorting out misinformation," https://healthfeedback.org/, 2024, accessed: 2024.

[21] Snopes Media Group, "Snopes: The definitive fact-checking site," https://www.snopes.com/, 2024, accessed: 2024.

[22] Agence France-Presse, "AFP fact check," https://factcheck.afp.com/, 2024, accessed: 2024.

[23] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," *arXiv preprint arXiv:1606.05250*, 2016.

[24] M. Abdin *et al.*, "Phi-3 technical report: A highly capable language model locally on your phone," *arXiv preprint arXiv:2404.14219*, 2024.

[25] H. Touvron *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[26] Mistral AI, "Mistral-7b-instruct-v0.2," https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2, 2024, instruction-tuned large language model.

[27] Meta AI, "The llama 3 herd of models," https://ai.meta.com/blog/meta-llama-3/, 2024, includes LLaMA-3-8B-Instruct.

[28] Y. Xiao, Z. Tang, P. Wei, C. Liu, and L. Lin, "Masked images are counterfactual samples for robust fine-tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 301–20 310.

[29] A. Farooq, S. Raza, M. N. Karim, H. Iqbal, A. V. Vasilakos, and C. Emmanouilidis, "Evaluating and regulating agentic ai: A study of benchmarks, metrics, and regulation," *Metrics, and Regulation*, 2025.