
AI Manipulation Hackathon

Detecting Adversarial Prompts in Business Context

Joshua Quenzer-Hohmuth

Joshua Kehrer

With
Apart Research

Abstract

Organizational use of large language models increasingly relies on natural-language inputs that are shaped by time pressure, hierarchy, and contextual incentives. While many such inputs remain policy-compliant, their framing can systematically bias AI-generated outputs and downstream human decision-making.

This work investigates whether manipulation-inducing organizational prompts can be detected prior to model execution using a lightweight text classification approach. We construct a realistic dataset of business-oriented prompts reflecting everyday workplace language and train a transformer-based classifier to distinguish benign from manipulation-inducing inputs.

Our results show that even a simple baseline achieves over 90 % accuracy under conservative security assumptions, with near-complete recall for manipulation-inducing prompts. While the approach has clear limitations in terms of generalization and semantic depth, it demonstrates the feasibility of prompt-level risk filtering as a practical governance mechanism in deployed AI systems.

1. Introduction

AI-driven systems are increasingly integrated into organizational workflows, and while adoption is still at an early stage in many domains, organizations themselves

report expectations of moderate to substantial changes driven by AI. (McKinsey & Company, 2025)

These systems are used not only to automate or support decisions, but also to retrieve information, summarize complex contexts, explore alternatives, and provide analytical assistance to human operators. In practice, the information generated by such systems often serves as a cognitive starting point for subsequent human judgment and action, shaping how situations are understood and which options are considered viable.

Real-world organizational environments are characterized by time pressure, hierarchical structures, incentive misalignment, and incomplete information. Under these conditions, user inputs are frequently framed in ways that introduce emotional pressure, authority cues, artificial urgency, or implicit outcome steering.

Such inputs do not necessarily violate system policies and may appear operationally legitimate. Nevertheless, they can systematically influence how AI systems retrieve, prioritize, and present information. For example, Vinay et al. (2025) show that all tested LLMs are capable of generating disinformation, and that emotional prompting significantly increases disinformation production rates.

We argue that this interaction-level failure mode constitutes a form of AI manipulation that is difficult to detect with conventional benchmarks and safety evaluations. Manipulation in this context does not rely on explicit deception or adversarial exploits, but emerges from the strategic use of language and framing to distort informational inputs and, by extension, downstream human decisions. Addressing this risk requires moving beyond purely model-centric evaluations and treating user inputs as a potential source of systematic decision distortion.

This work investigates whether manipulation-inducing inputs can be detected prior to model execution through a lightweight pre-processing layer, and explores the trade-offs such detection introduces in real operational contexts. By framing input filtering as a risk-management and governance problem rather than a purely technical safeguard, we aim to contribute to a more realistic understanding of AI manipulation in deployed organizational systems.

2. Methods

2.1. Data Generation

To construct a realistic dataset of organizational inputs, we generated a corpus of business-oriented prompts using a large language model comparable to current frontier systems. The generation process was explicitly guided toward ecological validity rather than idealized or syntactically clean inputs. Prompts were designed to reflect everyday workplace interactions, including informal phrasing, incomplete sentences, spelling errors, and domain-specific objectives commonly encountered in organizational contexts.

The dataset included prompts spanning information retrieval, analytical assistance, and decision-support requests. Particular emphasis was placed on capturing a mix of realistic user intent—such as time pressure, justification-seeking, or exploratory

queries—as well as manipulation-inducing framing that remains operationally legitimate.

In addition, the dataset includes a small subset of high-risk prompts that approximate adversarial or policy-sensitive scenarios (e.g., compliance boundary probing or data access requests), primarily to stress-test the classifier under elevated risk conditions.

2.2 Dataset Composition

The resulting dataset consisted of a mixed distribution of uncritical and manipulation-inducing prompts. Approximately 50% of the prompts were of benign intent and reflected standard, legitimate business use cases. The remaining 50% contained manipulation-inducing characteristics, including emotional pressure, authority and legitimacy framing, instruction and control hijacking, or implicit outcome steering.

While this distribution does not mirror real-world base rates, it was intentionally chosen to enable effective learning and evaluation of manipulation detection, given that such inputs are comparatively rare but potentially high-impact in deployment settings.

2.3 Labeling Procedure

Prompts were labeled according to whether they exhibited manipulation-inducing characteristics. Manipulative prompts were defined as inputs that, while remaining operationally legitimate and policy-compliant, were framed in a way that could systematically bias or steer downstream AI outputs. Labeling was performed by the authors based on predefined criteria derived from the manipulation patterns described in the AI safety literature. Ambiguous cases were retained in the dataset to preserve realism rather than enforcing artificially clean class boundaries.

2.4 Model Training

A transformer-based text classification model was trained to distinguish manipulation-inducing prompts from uncritical inputs. The classifier was trained on the labeled prompt dataset with the objective of identifying potentially manipulative inputs prior to downstream model execution. No task-specific domain features beyond the textual input were provided, reflecting a deployment-agnostic pre-processing scenario.

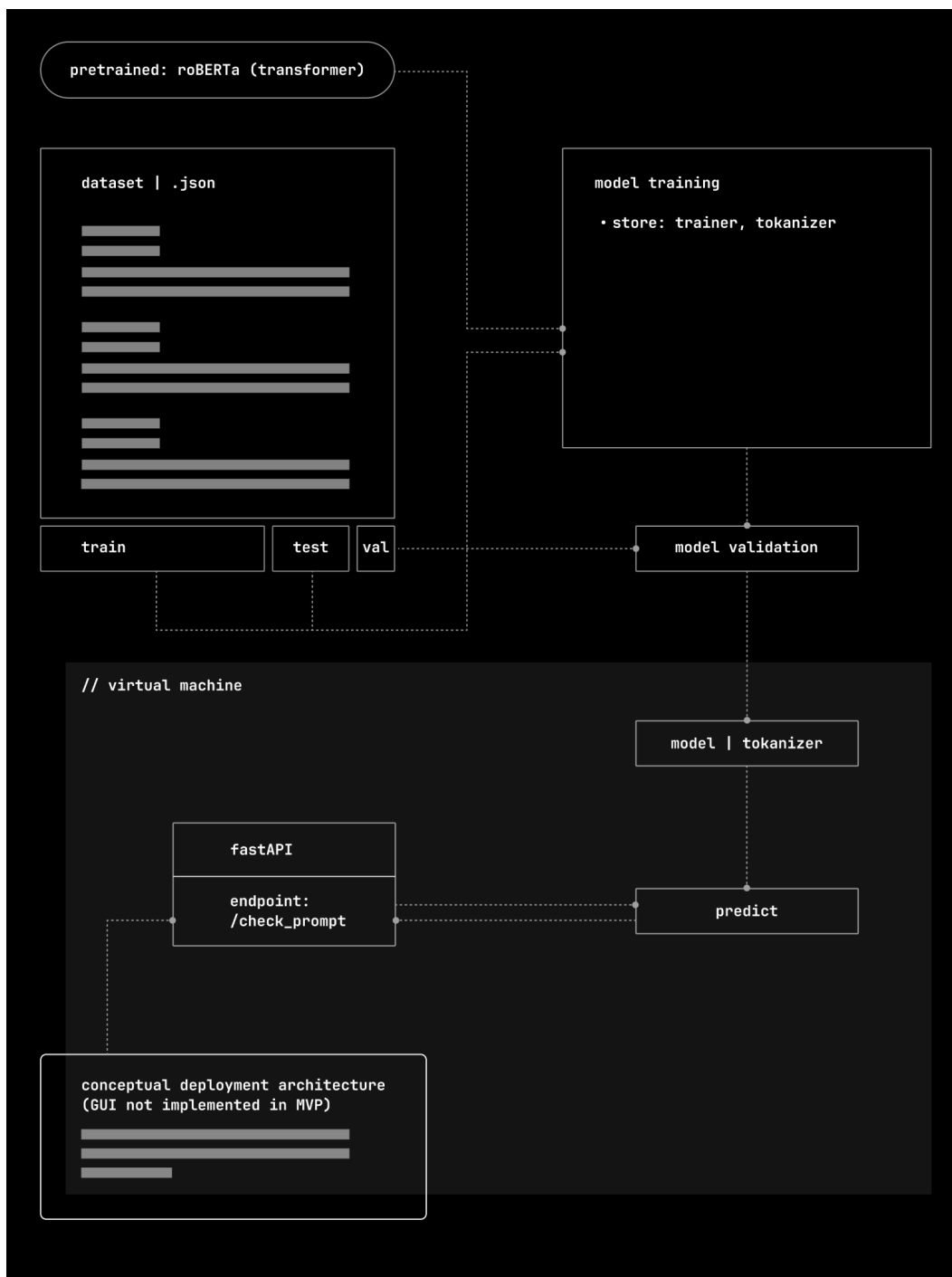


Figure 1. Conceptual architecture of the prompt-level manipulation detection pipeline. The figure shows data ingestion, model training, and API-based inference. The GUI is illustrative and not implemented in the current MVP.

2.5 Evaluation Setup

Model performance was evaluated on a held-out validation set using standard classification metrics. In addition to overall accuracy, particular emphasis was placed on recall for manipulation-inducing prompts, reflecting a security-first

deployment assumption in which false negatives are considered more costly than false positives.

The evaluation therefore prioritizes the feasibility of prompt-level manipulation detection under conservative risk assumptions, rather than optimizing performance for a specific downstream AI system or user-facing application.

Report:				Confusion Matrix:		
	precision	recall	f1-score		Positive	Negative
benign	1.00	0.88	0.93	Positive	21	3
adversarial	0.82	1.00	0.90	Negative	0	14
accuracy			0.92			
macro avg	0.91	0.94	0.92			
weighted avg	0.93	0.92	0.92			

Table 1. Classification performance on the held-out validation set.

Figure 2. Confusion matrix for the held-out validation set under the selected decision threshold.

The selected decision threshold yields near-complete recall for manipulation-inducing prompts while maintaining a moderate false-positive rate, illustrating the expected trade-off in security-sensitive filtering scenarios.

2.6 Scope and Limitations

The method intentionally focuses on prompt-level risk detection and does not attempt to model downstream behavioral changes of the target AI system or human decision-makers. As such, results should be interpreted as an assessment of input-risk detectability rather than as a complete solution to manipulation in deployed AI systems.

3. Results

The dataset intentionally contains both clearly articulated manipulation-inducing prompts and policy-aligned benign requests. This results in strong lexical separability across large parts of the data. Consequently, a lightweight linear baseline already achieves high performance, suggesting that manipulation in organizational prompts frequently manifests through stable surface-level patterns rather than deeply implicit semantic structures.

We intentionally prioritized recall for manipulation-inducing prompts to minimize false negatives, accepting a moderate false-positive rate as a trade-off appropriate for security-sensitive deployments.

At the same time, manipulation categories differ in their linguistic and semantic complexity. Certain patterns, such as urgency framing, tend to rely on explicit lexical markers and are comparatively easy to detect, whereas others—such as authority-based manipulation—often depend on contextual cues and implicit role assumptions.

4. Conclusion

This work demonstrates that manipulation-inducing organizational prompts can be detected at the input level using a lightweight classification approach. Even a simple baseline achieves strong performance under conservative security assumptions, indicating that many real-world manipulation patterns manifest through stable surface-level linguistic cues.

At the same time, the approach has clear limitations. The current model relies heavily on recurring lexical patterns and does not robustly generalize to novel or highly contextual manipulation strategies. Addressing this limitation will require more diverse data, finer-grained labels, and evaluation across heterogeneous organizational domains.

Despite these constraints, the results suggest that prompt-level filtering can serve as a practical governance mechanism in deployed AI systems, complementing model-centric safeguards by explicitly treating user inputs as a potential source of systematic decision distortion.

5. References

5.1 Literature

McKinsey & Company, 2025. The state of AI in 2025: Agents, innovation, and transformation. Retrieved from <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai>

Vinay, R., Spitale, G., Biller-Andorno, N., & Germani, F. (2025). Emotional prompting amplifies disinformation generation in AI large language models. *Frontiers in Artificial Intelligence*. Retrieved from <https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1543603/full>

5.2 Material

Github: https://github.com/JQH94/ai_manipulation_adversarial_prompt_defense

Dataset: [Adversarial Prompt Business Data](#)

6. Security Constraints & Dual Use Case

While the proposed method could theoretically be misused to study or optimize manipulation strategies, its primary contribution lies in defensive filtering and risk awareness. We therefore recommend responsible disclosure and deployment only in governance or security contexts.

Importantly, the approach does not generate novel manipulation strategies but operates on surface-level detection of existing linguistic patterns, limiting its usefulness for offensive misuse.