RogueGPT as Agentic Infrastructure: Seeding Cross-Lingual Misinformation Simulations and Red-Teaming Social Media Agents

Alexander Loth [1], Martin Kappes [1], and Marc-Oliver Pahl [2]

Abstract: The transition from static fake news generation to agentic disinformation campaigns represents a notable shift in online information dynamics. While Large Language Models (LLMs) have lowered the barrier for content creation, existing research often relies on monolithic or ad-hoc generation methods that do not capture the nuance required for Agent-Based Simulations (ABS) of misinformation diffusion. In this paper, we introduce **RogueGPT**, a modular, multi-model system designed to support misinformation diffusion studies. Unlike previous single-model solutions (e.g., GROVER), RogueGPT uses an extensible architecture supporting current models (GPT-4, LLaMA-3, Mistral, DeepSeek-R1) to generate controllable, parameter-rich news artifacts. We show how RogueGPT connects static text generation with dynamic agentic simulation, offering researchers control over persona-based stylistic attributes, narrative framing, and cross-lingual generation. Complementing this, **JudgeGPT** provides systematic human perception data collection, enabling validation of generated content against human authenticity judgments. By standardizing the "seeding" phase of misinformation simulations, RogueGPT supports benchmarking of detection systems against diverse content. We evaluate the system's effectiveness in generating content variants for testing propagation networks and multi-agent debate defense mechanisms, demonstrating its utility as a component of misinformation research infrastructure.

Keywords: Generative AI, Large Language Models, Agent-Based Simulation, Misinformation, Disinformation, Red-Teaming, Cross-Lingual, Human Perception, Research Infrastructure

1 Introduction

The proliferation of large language models (LLMs) has changed the information landscape, creating a dual-use paradigm where advanced tools facilitate both the detection and the sophisticated generation of harmful content [Sh24]. This technological change poses challenges for online safety, as LLMs can produce convincing misinformation that affects democratic processes, public health communication, and societal cohesion [LKP24], with collateral effects that ripple across digital ecosystems [LKP26d]. Recent events, including AI-generated content during electoral campaigns and fabricated news about geopolitical conflicts like the Israel-Hamas war [SJM24], underscore the need for research infrastructure to understand and counter this phenomenon.

[1] Frankfurt University of Applied Sciences, Faculty of Computer Science and Engineering, Nibelungenplatz 1, 60318 Frankfurt am Main, Germany,
alexander.loth@stud.fra-uas.de, https://orcid.org/0009-0003-9327-6865;
kappes@fra-uas.de, https://orcid.org/0000-0002-8768-8359

[2] IMT Atlantique, UMR IRISA, Chaire Cyber CNI, Brest, France,
marc-oliver.pahl@imt-atlantique.fr, https://orcid.org/0000-0001-5241-3809

While progress has been made in automated fake news detection [Ma23; SSM21] and harmful content identification [Ma19; St19], a gap exists in understanding the human dimension of this problem. As highlighted in our prior work [LKP24; Lo25], there is a need for empirical data on how humans perceive AI-generated content when the lines between human and machine-generated text are increasingly blurred. Current detection systems achieve high accuracy on benchmark datasets [SDM21; SN20], yet we lack systematic understanding of which factors make AI-generated misinformation convincing to human readers.

Core Research Question: How can we provide researchers with flexible tools to (1) generate diverse synthetic misinformation under controlled conditions, and (2) systematically collect human perception data to understand what makes AI-generated content believable?

We present **a system/demonstration contribution**: two complementary open-source tools forming a research pipeline, building on our prior work on synthetic news generation and human perception studies [LKP24; LKP26e; LKP26g; Lo25]. Rather than reporting a new empirical study, this paper describes the design, implemented functionality, and intended use of research infrastructure. Concretely, the infrastructure makes new research questions tractable: it lets researchers vary generation parameters as independent variables and link them to human authenticity judgments as dependent measures, enabling controlled, reproducible experiments on believability that ad-hoc prompting cannot support. The tools provide agentic infrastructure for seeding cross-lingual misinformation simulations and red-teaming social media agents:

- **RogueGPT**[3]: A multi-model stimulus generation engine integrating GPT-4, LLaMA, Mistral, and DeepSeek through a configurable JSON architecture. RogueGPT enables researchers to generate news fragments under precisely controlled experimental conditions—varying model, style (e.g., NYT, tabloid), format (tweet, article), and language—with full provenance tracking via MongoDB.
- **JudgeGPT**[4]: An empirical research platform where participants evaluate news authenticity through an interactive survey interface. JudgeGPT collects human judgments on perceived authenticity (real vs. fake) and origin (human vs. AI), creating a rich dataset linking generation parameters to quantitative perception scores.

Together, these tools enable a closed-loop experimental workflow: RogueGPT generates controlled stimuli with full metadata, JudgeGPT collects human perception data, and researchers can analyze which factors (model, style, topic) influence believability. This addresses the identified need for cross-platform, multilingual research infrastructure [SM24; SSM24].

Our contributions are:

[3] github.com/aloth/RogueGPT
[4] github.com/aloth/JudgeGPT

1. **Research Pipeline:** Two open-source tools providing end-to-end infrastructure from controlled stimulus generation to human data collection for studying AI-generated misinformation perception.
2. **Multi-Model Architecture:** Support for current LLMs (GPT-4, LLaMA, Mistral, DeepSeek) with configurable prompt engineering and provenance tracking.
3. **Human Perception Focus:** Unlike purely automated approaches, our pipeline captures human authenticity judgments, addressing the gap in understanding how people evaluate AI-generated content.
4. **Comparative Analysis:** A feature-level comparison with existing tools (GROVER, FACTGEN, FakeGPT) along explicitly defined dimensions (Section 4), clarifying where our design differs in scope rather than asserting general superiority.

Intended users and workflows. The infrastructure targets four overlapping communities, with distinct entry points: (i) *misinformation researchers* who need controlled stimuli and linked perception data; (ii) *red-teamers* who generate adversarial content variants to stress-test detectors; (iii) *agent-based simulation researchers* who need reproducible seed content for diffusion models; and (iv) *human-factors researchers* studying perception. A representative workflow: a researcher configures RogueGPT to produce style- and language-varied fragments, deploys them through JudgeGPT, and exports parameter-linked judgments for analysis.

This work aligns with the recognition that combating AI-generated misinformation requires interdisciplinary frameworks integrating technological capabilities with human-centered research [Ma21; Sh24]. By providing accessible tools, we aim to support research toward understanding and countering the diffusion of harmful content online.

The paper is structured as follows: Section 2 reviews related work on fake news generation and detection. Section 3 presents the RogueGPT and JudgeGPT architectures. Section 4 compares our tools with existing solutions. Section 6 summarizes contributions and outlines future directions.

2 Related Work

The study of AI-generated misinformation spans generation, detection, and human perception. In the context of LLMs and agents, research has expanded from isolated classification tasks to interconnected systems of generative agents, adversarial testing, and multimodal verification. We review three research clusters—Agent-Based Simulation, Agentic Red-Teaming, and Multimodal Verification—and describe how our tools support this research area.

2.1 Foundational Detection and Shared Tasks

The HASOC series [Ma19; Ma21; Ma23] established benchmarks for hate speech detection across Indo-European languages, while CheckThat! labs [SSM21] advanced fact-checking methodologies. COVID-19 accelerated health misinformation research [SDM21], with the FakeCovid dataset [SN20] highlighting the global, cross-lingual nature of misinformation diffusion. The CT-FAN dataset [Ma24] and FakeClaim [SJM24] extended this to multi-domain and conflict-related contexts. Cross-platform studies [SM24] reveal that harmful content migrates across linguistic and platform boundaries, underscoring the need for infrastructure that operates across multiple contexts.

2.2 Generative Approaches: From GROVER to LLMs

GROVER [Ze19] pioneered controllable fake news generation, conditioning text on metadata like headlines and publication dates. FACTGEN [Sh21] integrated external knowledge bases for factual consistency, while FakeGPT [HS24] explored prompt-based generation through ChatGPT. The release of open-weight models like LLaMA [To23] has democratized access to capable language models, enabling researchers to conduct reproducible experiments across model families. However, these foundational works—which provide important historical context—lack the multi-model flexibility and interactive interfaces required for contemporary research. As the field shifts from single-model systems to multi-agent architectures, new infrastructure is required to bridge generation, simulation, and evaluation.

2.3 Agent-Based Simulation of Misinformation

A notable development in computational misinformation research is LLM-powered Agent-Based Simulation (ABS). Vosoughi et al. [VRA18] established empirically that false news spreads faster and reaches more people than true news on social media—a finding that motivates the need for controlled generation tools to study these dynamics. Unlike traditional Agent-Based Modeling, where agents follow heuristic rules (e.g., "if neighbor shares, I share"), LLM-based ABS allows agents to process semantic content, exhibit cognitive biases, and generate novel text [Zh25]. Chuang et al. [Ch24] show that LLM agents can model confirmation bias and opinion dynamics, supporting the psychological plausibility of these simulations.

The Auditor-Node Framework. Maurya et al. [Ma25] introduce an "Auditor-Node" framework where LLM agents conditioned with specific personas ("Biased Individual," "Sensation Seeker") receive, rewrite, and propagate news articles. Their "Misinformation Index" tracks factual degradation as content traverses networks. While focusing on propagation mechanics, their reliance on standard LLM prompts for seed content limits experimental control. RogueGPT addresses this gap by serving as the "genesis engine" for such simulations,

offering granular control over initial rhetorical strategies. By standardizing seed content generation, researchers can isolate how initial stylistic choices influence subsequent propagation dynamics.

Generative Agents as Social Proxies. Park et al. [Pa23] introduced the foundational "Generative Agents" framework, demonstrating that LLM-powered agents can exhibit emergent social behaviors in simulated environments. Building on this paradigm, MOSAIC [Li25] demonstrates that generative agents can simulate granular user behaviors—liking, sharing, flagging content—with patterns that align statistically with human behavior. This validates LLM agents as proxies for social science research and intensifies the demand for diverse, high-fidelity synthetic stimuli. RogueGPT serves as the "Red Team" generator creating adversarial content to train and evaluate the "Blue Team" moderators in frameworks like MOSAIC. As Liu et al. demonstrate the validity of LLM agents as social proxies, the demand for methodologically controlled input content grows.

Semantic Diffusion. Zhang et al. [Zh25] contrast traditional probabilistic models (e.g., Independent Cascade) with "Semantic Diffusion," where influence is determined by message content and agent persona rather than network topology alone. In probabilistic models, text style is irrelevant; in semantic models, style ("Urgent," "Emotional," "Analytical") determines virality. RogueGPT's parameterization of style and tone enables rigorous empirical study of Semantic Diffusion dynamics, allowing researchers to isolate specific linguistic features that drive propagation.

2.4 Agentic Red-Teaming and Safety

The "Countering Misinformation" theme of contemporary research is increasingly dominated by Red Teaming—attacking AI systems to identify vulnerabilities. Weidinger et al. [We22] provide a comprehensive taxonomy of risks posed by language models, including misinformation and disinformation as key harm categories. Goldstein et al. [Go23] analyze how generative models enable automated influence operations, identifying specific threat vectors that infrastructure like RogueGPT can help researchers study and defend against. Perez et al. [Pe22] established the paradigm of using language models to red-team other language models, demonstrating that automated adversarial testing can uncover vulnerabilities at scale. By 2025, this evolved from manual "jailbreaking" to automated, agentic attack workflows [Pa24].

Dynamic Benchmark Generation. AgentHarm [An24a] establishes that "harm" in agentic contexts differs fundamentally from chatbot harm: agentic harm involves *doing* something (autonomously creating and spreading disinformation campaigns) rather than merely *saying* something offensive. RogueGPT provides the infrastructure for AgentHarm-style evaluations, automating generation of harmful content that agents must refuse. While static benchmarks provide vulnerability snapshots, our tools offer dynamic generation capability, allowing

safety researchers to continuously evolve test sophistication to match improving defense capabilities.

Compositional Attacks. Sun et al. [Su25] introduce "Composition of Principles" (CoP), where attacker agents combine multiple strategies (role-playing + logical reasoning + payload injection) to bypass defenses. RogueGPT's architecture, which separates prompt templates (strategy) from model execution, mirrors this compositional approach. Researchers can mix-and-match persona templates with rhetorical strategies, enabling systematic exploration of the "attack surface" of misinformation detection systems.

2.5 Detection, Verification, and Multimodality

Multi-Agent Debate for Verification. MAD-Sherlock [La25] represents state-of-the-art detection through multi-agent "jury" debate on content veracity. Testing such systems requires content sophisticated enough to provoke genuine debate—simple fake news is insufficient. RogueGPT's integration with advanced models (GPT-4, DeepSeek-R1) enables generation of "high-perplexity" misinformation: content that is logically consistent but factually false, suited for stress-testing debate architectures.

Evidence-Aware Detection. Bukke et al. [Bu25] propose an agentic multi-persona framework for evidence-aware misinformation detection, where specialized agents collaborate to verify claims. Such systems require training data that mimics the sophistication of real-world disinformation—precisely what RogueGPT's persona-driven generation provides.

Multimodal and Multilingual Infrastructure. Acknowledging the critical need for multilingual resources highlighted by Mandl et al. [Ma19; Ma24], RogueGPT integrates translation APIs for cross-lingual generation. Ji et al. [Ji23] provide a comprehensive taxonomy of hallucination in natural language generation, categorizing the failure modes that plague LLM outputs—a framework directly applicable to understanding how generated misinformation deviates from factual grounding. Jaiswal et al. [JLS25] extend this to multimodal contexts, demonstrating the growing challenge of hallucination in vision-language models. Future iterations will align with these multimodal frameworks by integrating image-generation endpoints, enabling creation of synthetic text-image pairs for comprehensive misinformation research.

2.6 Human Perception of AI-Generated Content

Beyond automated detection, understanding human perception remains critical. Studies on offensive language annotation [St19] reveal substantial inter-annotator disagreement reflecting content evaluation's subjective nature. Our prior work [LKP24; Lo25] identified the gap between algorithmic capabilities and human susceptibility. Complementary studies of *expert* perception show that even specialists disagree on how to counter AI-generated

disinformation and diverge by domain such as health versus politics [LKP26c; Lo26c], while causal analyses trace how repeated exposure erodes the human truth-default and drives susceptibility to foundation-model hallucinations in the wild [LKP26b]. Subsequent empirical studies validated key findings from the disinformation literature, revealing that LLMs achieve near-human mimicry (perception scores 0.46–0.50) and that demographic predictors show weaker effects for AI-generated content than human-written disinformation. Large-scale studies examining how demographics affect authenticity judgments demonstrate human-centered evaluation's importance.

2.7 Research Gaps Addressed

Existing tools exhibit limitations our work addresses: (1) single-model architectures limiting comparative studies across LLM families, (2) command-line interfaces reducing accessibility, (3) lack of integrated human perception data collection, (4) absence of controlled experimental workflows linking generation parameters to human judgments, (5) no support for seeding agent-based simulations with methodologically controlled content, (6) limited capability for dynamic adversarial dataset generation, and (7) no infrastructure for compositional attack research. RogueGPT and JudgeGPT address these gaps, providing multi-model generation integrated with systematic human perception data collection, designed to serve as infrastructure for ABS frameworks, red-teaming methodologies, and multi-agent verification systems.

3 RogueGPT and JudgeGPT: Architecture and Design

We present two complementary open-source tools forming a research pipeline for studying human perception of AI-generated misinformation: RogueGPT [Lo26a] for controlled stimulus generation and JudgeGPT [Lo25] for systematic human data collection. In the context of Agent-Based Simulation (ABS) and red-teaming, these tools provide seed content for misinformation simulations [Ma25] and support benchmark generation for adversarial testing [An24a].

3.1 Research Pipeline Overview

The experimental workflow connects controlled generation to human judgment (Figure 1):

1. **Controlled Stimulus Generation (RogueGPT):** Researchers generate news fragments with precisely defined parameters: LLM model, writing style (e.g., NYT, tabloid), format (tweet, article), and language.
2. **Data Storage (MongoDB):** Each fragment is stored with complete metadata (generation parameters, timestamp, prompt used), enabling full provenance tracking.

3. **Human Data Collection (JudgeGPT):** Participants evaluate fragments via an interactive survey, rating perceived authenticity (real/fake) and origin (human/AI).
4. **Analysis:** Researchers analyze relationships between generation parameters and human perception scores.

3.2 RogueGPT: Model-Agnostic Agent Backend

RogueGPT enables systematic generation of news fragments under controlled experimental conditions, supporting both human perception studies and agent-based simulation research.

3.2.1 Heterogeneous Agent Simulation via Multi-Model Support

Social networks are inherently heterogeneous, populated by actors with varying levels of linguistic sophistication and reasoning capability. RogueGPT addresses this reality by functioning as a **Model-Agnostic Agent Backend**. By supporting a diverse array of inference engines—from reasoning-heavy models to efficient, high-throughput alternatives—the framework enables simulation of "heterogeneous agent networks" [Zh25]:

- **High-Capability Agents**: GPT-4, GPT-4-turbo, DeepSeek-R1 for sophisticated state-actor simulation
- **Reactive Agents**: LLaMA 2/3, Mistral 7B for modeling organic user behavior
- **Cloud/Local Flexibility**: Azure OpenAI, local inference endpoints for deployment versatility

Researchers can assign different backend models to different agent nodes, empirically testing how the "intelligence gap" between high-resource state actors and low-resource organic users influences propagation dynamics of false narratives. Model selection is managed through `prompt_engine.json`, enabling systematic comparison without code changes.

3.2.2 Persona Configuration Module

The core of the architecture is the **Persona Configuration Module**. Rather than simple keyword injection, this module utilizes a templated system prompt architecture to define distinct "Agent Personas." Drawing on the methodology of Maurya et al. [Ma25], these templates encapsulate not just the topic of the misinformation, but the cognitive bias and rhetorical strategy of the generating agent. The `prompt_engine.json` configuration defines:

- **Style Personas**: News outlet personas (NYT institutional voice, BILD sensationalist, hyper-partisan skeptic)

- **Language**: Cross-lingual generation (English, German, French, Spanish)
- **Model Assignment**: Backend LLM for generation

This allows researchers to generate not isolated articles, but coherent "narrative streams" consistent with specific persona identities—a prerequisite for high-fidelity social simulations.

3.2.3 Diffusion Artifact Control

The system's **Diffusion Artifact Control** enables generation of content optimized for specific layers of a propagation network. By strictly enforcing structural constraints, RogueGPT supports:

- **Microblogging Artifacts**: Character-limited tweets for rapid, shallow diffusion
- **Summary Artifacts**: Headlines and short articles for mid-layer propagation
- **Long-Form Artifacts**: Full articles with journalistic conventions for deep engagement

This enables generation of "cascading" disinformation campaigns. For example, simulations can be seeded with a long-form "Article" artifact, which downstream agents then summarize into "Tweet" artifacts, allowing researchers to track how semantic nuance is lost or amplified as messages transmute across formats [Li25].

3.2.4 Human-in-the-Loop Red Teaming Interface

To support advanced **Hybrid Red-Teaming** workflows [Su25], RogueGPT incorporates a dedicated Human-in-the-Loop (HITL) interface via Streamlit:

- Dynamic parameter selection with real-time prompt preview
- Streaming output display during generation
- Automated batch generation across all parameter combinations
- **HITL Refinement Module**: Manual entry allowing domain experts to intercept and refine agent-generated payloads

While automated generation provides scale, it often lacks the subtle contextual awareness required to bypass frontier defense systems. The HITL module enables "Agentic Purple Teaming"—where human experts inject nuance or remove artifacts before content enters simulations, ensuring detection benchmarks are trained on the most potent, adversarially refined examples available [An24a].

3.3 JudgeGPT: Human Perception Data Collection Platform

JudgeGPT serves as the data collection platform where participants evaluate the authenticity of content generated by RogueGPT.

3.3.1 Survey Interface Design

The platform presents participants with news fragments and collects structured judgments:

- **Authenticity Rating**: Slider from "Definitely Fake" to "Definitely Real"
- **Origin Assessment**: Slider from "Definitely Human" to "Definitely AI"
- **Confidence Level**: Self-reported judgment confidence

3.3.2 Participant Engagement Features

To maximize data quality and participant retention, JudgeGPT incorporates:

- **Gamification**: Scores and performance feedback after each judgment
- **Personalized Results**: Comparison with aggregate participant performance
- **Multilingual Support**: Automatic browser language detection with manual override (EN, DE, FR, ES)
- **Demographic Collection**: Age, education, AI familiarity for analysis

3.3.3 URL-Based Experiment Configuration

Researchers can customize surveys via URL parameters:

- Language selection: `?language=de`
- Age filtering: `?min_age=18&max_age=35`
- Custom study identifiers for participant tracking

3.4 Data Model and Analysis Support

The shared MongoDB database links generation metadata to perception scores:

- **Fragments Collection**: Generated content with full provenance
- **Results Collection**: Human judgments linked to fragment IDs

- **Participants Collection**: Demographic and session data

The `data_analysis/` module provides tools for exporting data to CSV/JSON for statistical analysis, enabling researchers to investigate questions such as: Which LLM produces the most believable fake news? Does writing style affect perceived authenticity?

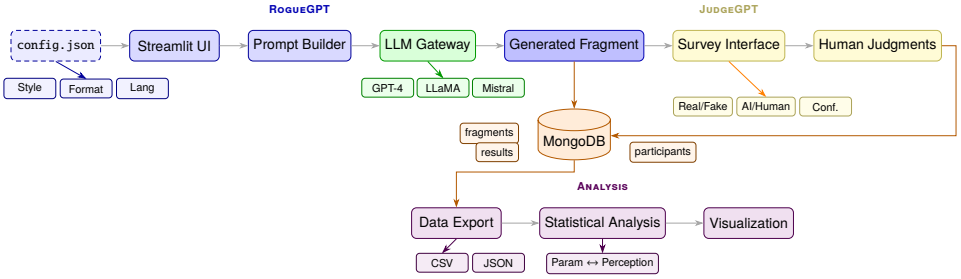

Fig. 1: Research pipeline: RogueGPT generates controlled stimuli via configurable LLM APIs (left); fragments with metadata are stored in MongoDB (center); JudgeGPT collects human authenticity judgments (right); analysis tools link generation parameters to perception scores (bottom).

4 Comparison with Existing Solutions

Table 1 compares our tools with notable fake news generation systems across key dimensions relevant to agent-based misinformation research. The comparison is *feature-scoped*: we assess design intent and implemented functionality along dimensions that matter for seeding simulations and red-teaming (model flexibility, persona/format control, bulk generation, human-in-the-loop support, provenance, multilinguality, openness). We do not claim superior generation quality or detection performance, which would require a separate empirical benchmark; the aim is to position our design relative to prior tools, several of which were built for different primary goals (e.g., GROVER for neural generation and detection, FACTGEN for fact-checking, FakeGPT for analysis).

4.1 Mapping to Simulation Requirements

Table 2 connects RogueGPT's features to the specific requirements identified in recent agent-based simulation literature, demonstrating how our infrastructure addresses emerging research needs.

4.2 Distinguishing Characteristics

Agentic Infrastructure. GROVER [Ze19], FACTGEN [Sh21], and FakeGPT [HS24] were each designed around a single model and a specific primary goal. RogueGPT instead targets

Feature	Ours	GROVER	FACTGEN	FakeGPT
Primary Paradigm				
Design Focus	Agentic	Neural Gen.	Fact-Check	Analysis
Model & Generation				
Model Backend	Multi	Single	Single	Single
Persona Config.	Templates	Metadata	Claim	Prompt
Format Control	Multi	Article	Article	Text
Simulation Readiness				
API/Bulk Gen.	✓	✗	✗	✗
HITL Support	Web UI	CLI	CLI	Chat
Research Infrastructure				
Human Data	✓	✗	✗	✗
Provenance	✓	✗	✗	✗
Multilingual	✓	✗	✗	Partial
Open Source	✓	✓	✗	✗

Tab. 1: Feature comparison with agentic dimensions. Ours = RogueGPT + JudgeGPT. Multi-model supports GPT-4, LLaMA, Mistral, DeepSeek.

simulation workflows: its multi-model support (GPT-4, LLaMA, Mistral, DeepSeek) is intended to enable heterogeneous agent networks where different nodes operate with varying capabilities [Zh25]. This is a difference in scope and design intent, not a claim of better output quality.

Human Perception Validation. JudgeGPT is designed to collect human authenticity judgments systematically, addressing a gap identified in misinformation research [Lo25]: understanding what makes AI-generated content believable to humans. We describe this as an *enabling capability* of the infrastructure rather than a completed validation; results from studies conducted with JudgeGPT are reported in our prior work and are out of scope for this system paper.

Research-Ready Pipeline. The integrated workflow from controlled generation through human data collection to analysis export provides end-to-end research infrastructure. MongoDB persistence with provenance tracking supports reproducibility—researchers can trace judgments back to generation parameters.

Accessibility. Interactive web interfaces lower barriers for researchers without extensive programming experience. This aligns with calls for making research infrastructure more accessible to interdisciplinary teams [Sh24].

4.3 Limitations

We describe implemented functionality rather than a completed empirical evaluation, and several limitations follow from this scope. (1) **Commercial API dependence:** generation

Simulation Requirement	Enabling Feature	Citation
Persona-based Propagation	Persona Configuration Module	Maurya [Ma25]
Opinion Dynamics	Confirmation Bias Modeling	Chuang [Ch24]
User Engagement Prediction	Diffusion Artifact Control (Tweets/Comments)	Liu [Li25]
Semantic Diffusion	Stylistic Parameterization (Tone/Sentiment)	Zhang [Zh25]
Multi-Step Agentic Harm	Automated Attack Workflows	Andriushchenko [An24a]
Automated Jailbreaking	Dynamic Red-Team Generation	Pavlova [Pa24]
Compositional Attacks	Modular Template Architecture	Sun [Su25]
Evidence-Aware Detection	Training Data for Multi-Persona Detectors	Bukke [Bu25]
Multi-Agent Debate	High-Perplexity Test Content	Lakara [La25]
Multilingual Verification	Cross-Lingual Generation	Mandl [Ma24]
Multimodal Testing	Extensible API (planned)	Jaiswal [JLS25]

Tab. 2: Mapping RogueGPT features to agent-based simulation and detection requirements from recent literature.

with proprietary models (e.g., GPT-4) incurs cost and availability constraints and ties reproducibility to third-party endpoints; local models (LLaMA, Mistral, DeepSeek) mitigate but do not remove this. (2) **Reproducibility across evolving foundation models:** model providers update or deprecate versions, so identical prompts can yield different outputs over time; provenance metadata records the model and configuration used, but exact regeneration is not guaranteed. (3) **Multilingual limits:** cross-lingual quality depends on the underlying model's coverage and degrades for lower-resource languages. (4) **Human data collection:** JudgeGPT requires participant recruitment and ethics/IRB approval, and gamified survey design needs validation for data quality. (5) **Evaluation scope:** this paper demonstrates architecture and functionality; systematic effectiveness studies (e.g., a user study quantifying believability across parameters) are future work. (6) **Misuse potential:** the generation component can produce misinformation-like content; we discuss mitigations in Section 6. Future work will address these through expanded local-model integration, larger and more diverse human studies, and systematic evaluation of survey design elements.

5 Future Directions

As misinformation research transitions from static analysis to dynamic agent-based simulation, RogueGPT should evolve from a standalone generator to an integrated component of agentic research infrastructure. We outline three development trajectories.

5.1 Integration with Multi-Agent Orchestrators

The immediate priority is direct integration with agent orchestration frameworks. RogueGPT's companion data layer already exposes a **Model Context Protocol (MCP)** server [An24b] with `ingest_fragment` and `retrieve_fragments` as standardized, agent-callable tools. This closes an autonomous ingest-retrieve loop today: an agent can generate content with any LLM, ingest the result through MCP with full provenance metadata, and retrieve fragments for subsequent human or automated evaluation—all without bespoke integration code. The natural next step is to expose generation itself as an MCP tool, alongside wrappers for LangChain, AutoGPT, and similar multi-agent systems, so that agents within a simulation can invoke RogueGPT to generate content dynamically in response to network feedback.

For example, when a previous post is fact-checked or debunked, a spreader agent could autonomously call RogueGPT to generate a contextually-aware rebuttal, adapting rhetorical strategy based on the debunking approach. This moves the tool from "Level 1" generation (human initiates) to "Level 2" generation (agent initiates autonomously), enabling study of feedback-driven disinformation dynamics [Ma25].

Concretely, we plan to implement:

- RESTful API with authentication for programmatic access
- Model Context Protocol (MCP) server exposing ingest and retrieval as agent-callable tools (live); generation as an MCP tool is the planned next step
- LangChain `Tool` wrapper for agent integration
- Callback hooks for logging agent-initiated generation events
- Rate limiting and quotas for simulation-scale workloads

5.2 Multimodal Disinformation Artifacts

While currently text-focused, RogueGPT's modular architecture is designed to support Multimodal Generative Agents. Building on the research on multimodal hallucination by Jaiswal et al. [JLS25], future work will integrate image-generation APIs (DALL-E 3, Midjourney, Stable Diffusion) to produce "Compound Misinformation Artifacts"—news articles paired with synthetic images.

This enables testing the robustness of multimodal verification systems against coordinated text-image disinformation. Specifically:

- **Semantic Consistency**: Generate images that "support" false textual claims, testing whether visual evidence increases believability
- **Cross-Modal Inconsistency**: Generate subtle mismatches between text and image to test detection of manipulation signals
- **Cascading Artifacts**: Track how multimodal content degrades as it propagates through text-only channels (e.g., image descriptions replacing images)

Integration with datasets like FakeClaim [SJM24] and TweetInfo [SSM24] will provide validation benchmarks for multimodal generation quality.

5.3 Sleeper Agent Simulation

An emerging safety concern involves "Sleeper Agents"—entities that exhibit normal behavior until triggered by specific narrative contexts [Hu24]. We plan to extend the Persona Configuration Module to support simulation of dormant disinformation networks.

The extended architecture will allow researchers to:

- Define **trigger phrases** or **narrative contexts** that activate latent personas
- Simulate networks where agents remain "benign" until coordinated activation signals appear
- Study the **cascade dynamics** when dormant networks activate simultaneously
- Test detection systems' ability to identify pre-positioned sleeper infrastructure

This capability addresses a gap in adversarial testing: most benchmarks assume continuous malicious behavior, while real-world influence operations often involve prolonged dormancy followed by coordinated activation around key events (elections, crises, etc.).

5.4 Cross-Lingual Propagation Studies

Building on the multilingual infrastructure needs highlighted by Mandl et al. [Ma24] and the cross-platform diffusion patterns identified by Shahi et al. [SM24], we will extend RogueGPT's cross-lingual capabilities:

- **Translation-Aware Generation**: Generate content designed for subsequent translation, testing how disinformation degrades or amplifies across language boundaries
- **Code-Switching Artifacts**: Produce content mixing languages for multilingual communities

- **Cultural Adaptation**: Persona templates that adapt rhetorical strategies to cultural contexts while preserving core disinformation payload

These extensions position RogueGPT as infrastructure for studying the global, cross-lingual nature of modern disinformation campaigns.

6 Conclusion and Future Work

This work presented RogueGPT and JudgeGPT as a system/demonstration contribution: research infrastructure for studying misinformation in the context of LLMs and agents. By moving beyond static generation approaches, RogueGPT provides infrastructure to seed agent-based simulations and red-teaming workflows. We described its design and demonstrated its flexibility across multiple models and its ability to generate persona-driven content for testing diffusion dynamics; systematic effectiveness studies remain future work.

Key Contributions:

1. **Agentic Infrastructure:** RogueGPT functions as a Model-Agnostic Agent Backend, enabling heterogeneous agent simulations where different nodes operate with varying linguistic sophistication [Zh25].
2. **Persona Configuration:** Drawing on the methodology of Maurya et al. [Ma25], our templated architecture supports persona-conditioned generation for social simulations.
3. **Diffusion Artifact Control:** Support for cascading content formats (articles → tweets) enables study of semantic degradation across propagation layers [Li25].
4. **Human Validation Layer:** JudgeGPT is designed to collect systematic perception data that links generation parameters to human authenticity judgments, an enabling capability for assessing believability in future simulation studies [Lo25].

Addressing Infrastructure Needs: Our comparison with existing tools highlights the need for dynamic, controllable generation that feeds into propagation simulations like MOSAIC [Li25] and LLM-AIDSim [Zh25]. RogueGPT addresses this need, offering a testbed for research into cross-lingual disinformation, the "Semantic Diffusion" dynamics described by Zhang et al., and behaviors of agentic social networks.

Future Directions: Building on our empirical validation studies [LKP24; Lo25], we plan extensions including: multimodal support for synthetic text-image pairs, aligning with Jaiswal et al.'s [JLS25] work on vision-language hallucination; integration with multi-agent debate architectures like MAD-Sherlock [La25]; and compositional attack generation following the CoP framework [Su25]. We also aim to expand cross-lingual capabilities, addressing the multilingual infrastructure needs highlighted by Mandl et al. [Ma24].

Supporting the Full Disinformation Lifecycle: While RogueGPT centers on the content-generation phase, the infrastructure is intended to support research across the AI-enabled

disinformation kill chain [LKP26e]. Reconnaissance and narrative design are addressed through persona and framing configuration; *generation* through multi-model, multilingual, multi-format output; *distribution and adaptation* through diffusion-artifact control (article-to-tweet cascades) that seeds propagation simulations such as MOSAIC [Li25]; and *detection and defense* through provenance-tracked stimuli and JudgeGPT judgments that feed red-teaming and multi-agent debate architectures [La25]. Downstream defenses can build on complementary components of our research program, including on-device cryptographic image provenance [Lo26b], domain-level credibility priors for pre-bunking [LKP26a]—itself exposed as an MCP server so verification agents can query source reliability as a standardized tool—and reproducible provenance workflows motivated by the expert-perceived verification crisis [LKP26f]. We make no claim to cover every phase completely; rather, the tools provide connective infrastructure that lets researchers study individual phases and their transitions under controlled conditions.

Ethical Considerations and Responsible Release: RogueGPT can generate misinformation-like content, so we treat misuse prevention as a first-class design concern rather than an afterthought. Our mitigations span four areas. (1) *Access control:* generation requires researcher-supplied API keys or locally hosted models, so the tool ships no ready-to-abuse hosted generation endpoint; the live JudgeGPT survey exposes only pre-vetted stimuli, not open generation. (2) *Labeling and provenance:* every generated artifact is tagged with model, configuration, and generation metadata in MongoDB, so synthetic content is traceable and distinguishable from authentic material by design, and outputs are intended for research contexts, not publication. (3) *Responsible release:* we publish documentation that states the intended research use, require ethics/IRB approval for human-subject studies, and orient the tooling toward detection, prebunking, and red-teaming rather than deployment. (4) *Residual risk:* we acknowledge that open-sourcing generation tooling carries dual-use risk; we judge the benefit to defensive research to outweigh it given that comparable generation capability is already widely accessible through general-purpose LLMs, and we invite community norms on gated access for higher-risk extensions (e.g., large-scale automated campaigns). We commit to revisiting the release posture as capabilities and misuse patterns evolve.

As the information ecosystem becomes increasingly populated by autonomous social media agents, the ability to simulate their output with precision becomes a practical research requirement. We open-source this research infrastructure to foster a collaborative approach to countering misinformation [An24a; Sh24].

Data and Code Availability

RogueGPT: github.com/aloth/RogueGPT
JudgeGPT: github.com/aloth/JudgeGPT
Live Survey: judgegpt.streamlit.app
Both repositories include documentation, example configurations, and data analysis tools.

References

[An24a] Andriushchenko, M. et al.: AgentHarm: A Benchmark for Measuring Harmfulness of LLM Agents. arXiv preprint arXiv:2410.09024, 2024.

[An24b] Anthropic: Model Context Protocol, https://modelcontextprotocol.io/, Open protocol for connecting AI assistants to tools and data sources, 2024.

[Bu25] Bukke, N. V. S. V. K. et al.: Agentic Multi-Persona Framework for Evidence-Aware Misinformation Detection. arXiv preprint arXiv:2502.18951, 2025.

[Ch24] Chuang, Y.-S. et al.: Simulating Opinion Dynamics with Networks of LLM-based Agents. arXiv preprint arXiv:2311.09618, 2024.

[Go23] Goldstein, J. A. et al.: Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. arXiv preprint arXiv:2301.04246, 2023.

[HS24] Huang, Y.; Sun, L.: FakeGPT: Fake News Generation, Explanation and Detection of Large Language Models, 2024, arXiv: 2310.05046 [id='cs.CL'].

[Hu24] Hubinger, E. et al.: Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training. arXiv preprint arXiv:2401.05566, 2024.

[Ji23] Ji, Z. et al.: Survey of Hallucination in Natural Language Generation. ACM Computing Surveys 55 (12), pp. 1–38, 2023.

[JLS25] Jaiswal, A. K.; Liu, H.; Shahi, G. K.: Multimodal Hallucination Detection and Mitigation in Large Vision-Language Models. arXiv preprint arXiv:2501.02114, 2025.

[La25] Lakara, K. et al.: MAD-Sherlock: Multi-Agent Debate for Visual Misinformation Detection. In: ICML 2025 Workshop on Collaborative and Federated Agentic Workflows. 2025.

[Li25] Liu, G. et al.: MOSAIC: Modeling Social AI for Content Dissemination and Regulation in Multi-Agent Simulations. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Pp. 6401–6428, 2025.

[LKP24] Loth, A.; Kappes, M.; Pahl, M.-O.: Blessing or curse? A survey on the Impact of Generative AI on Fake News, 2024, arXiv: 2404.03021 [cs.CL].

[LKP26a] Loth, A.; Kappes, M.; Pahl, M.-O.: CRED-1: An Open Multi-Signal Domain Credibility Dataset for Automated Pre-Bunking of Online Misinformation. Preprint available at SSRN, 2026, https://ssrn.com/abstract=6448466.

[LKP26b] Loth, A.; Kappes, M.; Pahl, M.-O.: Eroding the Truth-Default: A Causal Analysis of Human Susceptibility to Foundation Model Hallucinations and Disinformation in the Wild. In: Companion Proceedings of the ACM Web Conference 2026 (WWW '26 Companion). ACM, Dubai, United Arab Emirates, pp. 1125–1132, 2026, https://arxiv.org/abs/2601.22871.

[LKP26c] Loth, A.; Kappes, M.; Pahl, M.-O.: Experts Disagree on How to Fight AI Disinformation, but Agree That Health and Politics Need Different Solutions. Harvard Kennedy School Misinformation Review 7 (4), 2026.

[LKP26d] Loth, A.; Kappes, M.; Pahl, M.-O.: Industrialized Deception: The Collateral Effects of LLM-Generated Misinformation on Digital Ecosystems. In: Companion Proceedings of the ACM Web Conference 2026 (WWW '26 Companion). ACM, Dubai, United Arab Emirates, pp. 295–302, 2026, https://arxiv.org/abs/2601.21963.

[LKP26e] Loth, A.; Kappes, M.; Pahl, M.-O.: Interrupting the Chain: An Empirical Analysis of the AI-Driven Disinformation Kill Chain Through Human Perception Data. In: Proceedings of the ACM Web Conference 2026. WWW '26, ACM, 2026.

[LKP26f] Loth, A.; Kappes, M.; Pahl, M.-O.: The Verification Crisis: Expert Perceptions of GenAI Disinformation and the Case for Reproducible Provenance. In: Companion Proceedings of the ACM Web Conference 2026 (WWW '26 Companion). ACM, Dubai, United Arab Emirates, pp. 980–988, 2026, https://arxiv.org/abs/2602.02100.

[LKP26g] Loth, A.; Kappes, M.; Pahl, M.-O.: Verifying the Illusion: Presentation and Reproduction of Key Results from 5 Years of Empirical Research on Automation-Fueled Disinformation. In: Proceedings of the 18th ACM Web Science Conference 2026. WebSci '26, ACM, 2026.

[Lo25] Loth, A. et al.: Deceived by the Machine? Generative AI and LLMs for FakeNews: JudgeGPT—a continuous study sensibilizing and prebunking humans. In: IEEE Transactions on Human-Machine Systems. Accepted, 2025.

[Lo26a] Loth, A. et al.: Generative AI and LLMs as Game Changer for FakeNews? In: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems. CHI '26, ACM, 2026.

[Lo26b] Loth, A. et al.: Origin Lens: Reclaiming Trust on the AI-Mediated Web Through On-Device Image Provenance Verification. In: 18th ACM Web Science Conference (WebSci Companion '26), May 26–29, 2026, Braunschweig, Germany (WebSci '26). ACM, Braunschweig, Germany, pp. 1–5, 2026.

[Lo26c] Loth, A. et al.: When Experts Disagree: Mapping Consensus and Conflict in Expert Assessments of AI-Generated Disinformation. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES 2026). Accepted. To appear., Malmö, Sweden, 2026.

[Ma19] Mandl, T. et al.: Overview of the HASOC Track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages. In: Proceedings of the 11th Forum for Information Retrieval Evaluation (FIRE). ACM, pp. 14–17, 2019.

[Ma21] Mandl, T. et al.: Overview of the HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages. In: arXiv preprint arXiv:2112.09301. 2021.

[Ma23] Mandl, T. et al.: Overview of the HASOC Subtrack at FIRE 2022: Hate Speech and Offensive Content Identification. In: Proceedings of the 14th Forum for Information Retrieval Evaluation (FIRE). ACM, 2023.

[Ma24] Mandl, T. et al.: CT-FAN: A Dataset for Multi-domain Fact Checking and Claim Analysis. In: Proceedings of the Language Resources and Evaluation Conference (LREC). ELRA, 2024.

[Ma25] Maurya, R. G. et al.: Simulating Misinformation Propagation in Social Networks using Large Language Models. arXiv preprint arXiv:2511.10384, 2025.

[Pa23] Park, J. S. et al.: Generative Agents: Interactive Simulacra of Human Behavior. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST). ACM, pp. 1–22, 2023.

[Pa24] Pavlova, A. et al.: GOAT: A Global Optimization-Based Approach for Automated Multi-Turn Jailbreaking of Large Language Models. arXiv preprint arXiv:2410.01606, 2024.

[Pe22] Perez, E. et al.: Red Teaming Language Models with Language Models. arXiv preprint arXiv:2202.03286, 2022.

[SDM21] Shahi, G. K.; Dirkson, A.; Majchrzak, T. A.: An exploratory study of COVID-19 misinformation on Twitter. Online Social Networks and Media 22, p. 100104, 2021.

[Sh21] Shu, K. et al.: Fact-enhanced synthetic news generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35 15. 15, pp. 13825–13833, 2021.

[Sh24] Shahi, G. K. et al.: Report on the 1st Workshop on Diffusion of Harmful Content on Online Web (DHOW) at WebSci 2024. In: Companion Publication of the 16th ACM Web Science Conference. ACM, 2024.

[SJM24] Shahi, G. K.; Jaiswal, A. K.; Mandl, T.: FakeClaim: A Multiple Platform-driven Dataset for Identification of Fake News on 2023 Israel-Hamas War. In: European Conference on Information Retrieval (ECIR). Springer, pp. 66–74, 2024.

[SM24] Shahi, G. K.; Majchrzak, T. A.: Hate Speech Detection Using Cross-platform Social Media Data in English and German Language. In: arXiv preprint arXiv:2410.05287. 2024.

[SN20] Shahi, G. K.; Nandini, D.: FakeCovid–A Multilingual Cross-domain Fact Check News Dataset for COVID-19. arXiv preprint arXiv:2006.11343, 2020.

[SSM21] Shahi, G. K.; Struß, J. M.; Mandl, T.: Overview of the CLEF-2021 CheckThat! Lab: Task 3 on Fake News Detection. In: CEUR Workshop Proceedings. Vol. 2936, 2021.

[SSM24] Shahi, G. K.; Struß, J. M.; Mandl, T.: TweetInfo: An Event-based Multimodal Dataset for Detecting Misinformation on Online Social Networks. In: Proceedings of the ACM Web Science Conference. 2024.

[St19] Struß, J. M. et al.: Overview of GermEval Task 2, 2019: Shared Task on the Identification of Offensive Language. In: Proceedings of the 15th Conference on Natural Language Processing (KONVENS). Pp. 354–365, 2019.

[Su25] Sun, Z. et al.: CoP: Agentic Red-teaming for Large Language Models using Composition of Principles. In: Advances in Neural Information Processing Systems (NeurIPS). 2025.

[To23] Touvron, H. et al.: LLaMA: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971, 2023.

[VRA18] Vosoughi, S.; Roy, D.; Aral, S.: The spread of true and false news online. Science 359 (6380), pp. 1146–1151, 2018.

[We22] Weidinger, L. et al.: Taxonomy of Risks Posed by Language Models. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 214–229, 2022.

[Ze19] Zellers, R. et al.: Defending against neural fake news. Advances in neural information processing systems 32, 2019.

[Zh25] Zhang, L. et al.: LLM-AIDSim: LLM-Enhanced Agent-Based Influence Diffusion Simulation in Social Networks. Systems 13 (1), p. 29, 2025.