The Journal of Open Source Software

RogueGPT: A Controlled Stimulus Generation Framework for News Authenticity Research

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

1 Frankfurt University of Applied Sciences, Germany **2** IMT Atlantique, France

DOI: 10.21105/joss.11219

Software

- Review
- Repository
- Archive

Editor: Chris Vernon

Reviewers:

- @dosvk
- @prajwal-svm

Submitted: 13 June 2026
Published: 12 September 2026

Summary

RogueGPT is an open-source Python framework for the controlled, reproducible generation and curation of multilingual news fragments for AI authenticity research. It enables researchers to systematically produce synthetic news stimuli across a wide range of large language model (LLM) families, journalistic styles, languages, and content formats, while storing every fragment alongside complete generation provenance in a MongoDB corpus.

The framework follows a three-layer architecture that strictly separates the data logic (`core.py`) from user-facing interfaces: a Streamlit web application, a command-line interface (CLI), and a Model Context Protocol (MCP) server for AI-agent integration. All three interfaces share a single validation and normalisation layer, ensuring that every fragment — whether ingested manually, via automated generation scripts, or through an AI agent — conforms to the same schema. Each machine-generated fragment records the model identifier, the full prompt, the sampling parameters used by the batch generator, and the language, format and seed phrase where the interface supplies them. Fragments created through the web interface record model and prompt but not sampling settings, because that path delegates them to the provider defaults; the stored record therefore states what was fixed rather than implying that every generation setting is recoverable.

The current corpus contains 3,278 multilingual news fragments. Of these, 2,638 are machine-generated by 10 models across 6 providers (OpenAI, Google, Meta, Anthropic, Mistral, and Microsoft), covering four languages (English, German, French, and Spanish), three content formats (tweet, headline, and short article), and five journalistic styles per language. The remaining 640 fragments are human-sourced — both legitimate news and authentic fake-news material — and serve as experimental anchors for perception studies.

RogueGPT is the upstream stimulus generation component of a three-tool research pipeline: CRED-1 (Loth, 2026a) identifies unreliable news sources, RogueGPT generates controlled stimuli, and JudgeGPT delivers them to human participants for perception measurement.

Statement of Need

Empirical research on human and automated detection of AI-generated news requires stimuli that are diverse, systematically varied across experimental conditions, and fully reproducible. Existing corpora are typically static snapshots generated by a single model or a narrow set of models, lacking the fine-grained metadata required to isolate the effect of individual experimental variables (model family, version, prompt strategy, language, style) on detection outcomes (Kreps et al., 2022; Zellers et al., 2019).

RogueGPT addresses this gap in three ways:

1. **Systematic parameterisation.** Stimuli are generated by Cartesian expansion of a

declarative configuration (`prompt_engine.json`) that defines models, languages, styles, and formats. Every combination is reachable without manual intervention, enabling researchers to expand the corpus to new model releases in minutes by updating a single JSON file.

2. **Complete provenance.** Every fragment is stored with the exact model identifier, prompt text, ISO language code, content format, journalistic style, ingestion channel, sampling parameters (for newly generated fragments), and a UTC creation timestamp. This makes it possible to reconstruct the conditions under which any stimulus was produced and to filter the corpus along any experimental axis at query time via the CLI or MCP API.

3. **Living corpus design.** Unlike frozen datasets, the corpus is designed to grow. The MCP server exposes `ingest_fragment` and `retrieve_fragments` as tool-use endpoints, enabling AI agents to autonomously extend the corpus with new model outputs as part of larger automated pipelines. This is particularly important given the rapid pace of LLM releases: a framework that supports incremental corpus expansion without re-running entire generation pipelines is essential for longitudinal research.

The framework is part of a broader open-source research ecosystem for studying online misinformation. Upstream, the CRED-1 Domain Credibility Dataset (Loth, 2026a) provides the automated source selection that identifies unreliable domains whose content is scraped as human-authored fake news anchors. Downstream, JudgeGPT delivers RogueGPT-generated stimuli to human participants and records dual-axis perception judgements. Together, the three tools form an end-to-end pipeline from source identification through stimulus generation to quantitative perception measurement.

RogueGPT is already in active use in the following peer-reviewed publications. All four originate from the authors' own research programme, which reflects the stage of adoption rather than a claim of external uptake:

- *Industrialized Deception* (Loth, Kappes, et al., 2026b) analyses the systemic effects of LLM-generated misinformation on digital ecosystems, using stimuli generated by RogueGPT (ACM WWW '26).
- *Eroding the Truth-Default* (Loth, Kappes, et al., 2026a) reports human perception data from 154 participants who contributed 918 evaluations of stimuli produced by this framework (ACM WWW '26).
- *The Verification Crisis* (Loth, Kappes, et al., 2026c) presents expert survey findings on GenAI disinformation and reproducible provenance as part of the same research programme (ACM WWW '26).
- The foundational survey *Blessing or Curse?* (Loth et al., 2024) describes the methodology underpinning the generation pipeline.

The full corpus produced by RogueGPT is archived on Zenodo (concept DOI: 10.5281/zenodo.18703137, which always resolves to the current version) under restricted rather than open access. Access is granted to researchers at academic or non-profit institutions on request. The restriction has two reasons. The machine-generated fragments are released under CC BY 4.0, but the human-sourced fragments are excerpts of third-party news material that the depositor cannot license onward, and they are shared under the research exception for text and data mining, which limits sharing to researchers rather than the general public. The restriction also keeps a corpus of misinformation-like material from being distributed without an accountable requester. Every fragment records its own origin. The machine-generated rows carry `MachineModel` and `MachinePrompt`; the human-sourced rows are traceable to outlet and URL through the `HumanOutlet` and `HumanURL` columns, which the deposit carries alongside `IngestedVia` for every fragment. The companion JudgeGPT deposit (10.5281/zenodo.22226580) ships a byte-identical copy of the same corpus. The `Origin` column keeps the two parts separable and independently filterable.

Key Features

Multi-Model Support

The model registry in `prompt_engine.json` defines 47 identifiers across 11 providers as of 2026-02-23, a capability surface that is deliberately broader than any single corpus snapshot: registry entries are available for generation and do not imply that a model contributed to the released corpus. The naming convention (`provider_model-variant`) is provider-agnostic; adding a new model requires a single line change to the JSON file and no code modifications. At ingest time, the CLI and MCP server validate submitted model identifiers against the registry, with an optional `--lenient` flag for provisional use of unreleased models.

Structured Fragment Schema

Each fragment is stored with complete provenance metadata including model identifier, prompt text, ISO language code, content format, journalistic style, source outlet, and UTC timestamp. A shared validation layer enforces schema consistency across all three interfaces, ensuring that every fragment — regardless of ingestion path — carries the metadata required to reproduce the experimental design. Newly generated fragments additionally record their sampling parameters alongside the model identifier and prompt, and the validator reports a warning when that field is absent, so fragments ingested before this was introduced remain identifiable as such. Exact regeneration of a given output remains outside the framework's control in any case, since commercial APIs are non-deterministic and providers revise models behind stable identifiers. The full schema specification is documented in the repository.

Interfaces

RogueGPT provides three user-facing interfaces that share a single validation layer: a command-line interface for scripted ingestion and retrieval, a Streamlit web application for interactive use, and a Model Context Protocol (MCP) server that enables LLM-based agents to query the corpus and ingest new fragments programmatically. The MCP integration allows AI agents to autonomously extend the corpus as part of larger automated research pipelines.

Human Fragment Sourcing

Human-written fragments are sourced from two pools:

- **Legitimate news:** Opening paragraphs from established international outlets (BBC, The Guardian, Reuters, Tagesschau, FAZ, Le Monde, El País) via RSS feeds.
- **Fake news:** Articles scraped from domains classified as *fake*, *unreliable*, or *conspiracy* by the CRED-1 Domain Credibility Dataset (Loth, 2026a), which scores 2,672 domains using editorial classification, fact-check claims, web popularity, and domain age.

Both categories are ingested with full provenance metadata, providing experimental anchors that reflect authentic disinformation language rather than artificially constructed examples.

Software Design

RogueGPT is organised in three layers so that the data logic can be tested and reused independently of any interface.

`core.py` is the only module that talks to MongoDB. It owns schema validation, field normalisation, config loading, and the retrieval API. It has no UI dependencies, which is what makes the test suite able to cover it directly.

Three interfaces sit on top of `core.py` and share its validation path: a Streamlit web application (`app.py`) for interactive generation, a command-line interface (`cli.py`) for scripted and batch

work, and a Model Context Protocol server (`mcp_server.py`) that exposes `ingest_fragment` and `retrieve_fragments` as tools for AI-agent workflows. Because all three route through the same validator, a fragment ingested by an agent is subject to exactly the same schema and provenance requirements as one entered by hand.

Generation behaviour is not hard-coded. `prompt_engine.json` declares the model identifiers, prompt templates, languages, journalistic styles, and content formats, so extending the corpus to a newly released model is a configuration change rather than a code change. Every stored fragment carries the parameters that produced it, which is what allows a later analysis to filter by any single experimental variable.

State of the Field

Several tools and datasets address fake news generation and detection research, but none provide an active, extensible generation infrastructure with full provenance tracking.

GROVER (Zellers et al., 2019) generates and detects neural fake news but is limited to a single GPT-2-scale architecture and English text. The LIAR dataset (W. Y. Wang, 2017) provides 12,836 human-labelled claims across six fine-grained labels but offers no generation infrastructure. TruthfulQA (Lin et al., 2022) measures model truthfulness across 817 questions but targets factual consistency rather than controlled stimulus generation. Kreps et al. (2022) demonstrate that GPT-2-generated political news is perceived as credible as human-written content, establishing the need for systematic multi-model comparisons — but their work uses a single model without a reusable generation framework.

More recent efforts such as the M4 corpus (Y. Wang et al., 2024) collect multi-generator outputs across domains and languages but distribute them as static datasets without tooling for reproducible expansion. Similarly, MULTITuDE (Macko et al., 2023) covers 8 multilingual LLMs across 11 languages but is frozen at its release configuration. FakeNewsNet (Shu et al., 2020) provides a rich repository combining news content with social context but focuses on detection research with pre-collected articles rather than controlled stimulus generation.

RogueGPT differs from all of these in three respects: (1) it is an active generation infrastructure rather than a frozen benchmark, supporting incremental corpus expansion as new models are released; (2) it stores complete generation provenance per fragment, enabling fine-grained experimental filtering by model, language, style, and format; and (3) its MCP integration enables direct use within AI agent workflows; we are not aware of another research corpus management system that exposes such an interface.

Research Impact Statement

RogueGPT exists to measure how people and detectors respond to AI-generated news, and its design follows from that purpose. The framework produces experimental stimuli for controlled studies; it has no publishing, posting, or distribution path, and no scheduling, targeting, or audience-selection functionality. Fragments are written to a private MongoDB corpus and are read back by researchers.

The scientific value lies in control and provenance. Because every fragment records the model, prompt strategy, language, style, and format that produced it, effects can be attributed to a specific variable rather than to an undifferentiated notion of "AI text". This is what static, single-model corpora cannot support, and it is the property that downstream perception work depends on. RogueGPT is the upstream stage of a three-tool pipeline: CRED-1 (Loth, 2026a) identifies unreliable sources, RogueGPT generates controlled stimuli, and JudgeGPT collects human judgements that are linked back to the generation parameters. Results from this pipeline have been reported in peer-reviewed venues (Loth, Kappes, et al., 2026b, 2026a).

We recognise that a stimulus generator for misinformation research carries dual-use risk, and we address it through the design rather than through disclaimers alone. The framework generates short fragments for study delivery, not complete or distributable articles. It requires the researcher to supply their own API credentials, so generation remains attributable to an accountable party under the terms of the upstream providers, whose own usage policies continue to apply. Human data collection through JudgeGPT is subject to institutional ethics review. The corpus is archived on Zenodo under restricted access for academic use rather than released openly (Loth, 2026b). The software is licensed under GPL-3.0, so derivative work remains open to inspection.

The counterfactual matters here: the generative capability RogueGPT orchestrates is already available to anyone through the same commercial APIs it calls, and it adds none. What does not currently exist in the open is a reproducible, fully instrumented way for researchers to study that capability under controlled conditions. Withholding the measurement infrastructure would not reduce the availability of synthetic misinformation; it would leave the research community without a shared basis for studying it.

AI Usage Disclosure

The software generates its research stimuli by calling commercial large language model APIs; this is the documented function of the framework and is described throughout this paper.

For the development of the software and the preparation of this manuscript, Claude Opus 4.8 was used through GitHub Copilot in Visual Studio Code. In the software, it was used for code completion, refactoring suggestions, and drafting of docstrings. In the manuscript, it was used to generate the BibTeX bibliography from the authors' Zotero export. No AI system was used to design the study, to build or analyse the corpus, or to formulate the findings reported here.

All AI-assisted output has been reviewed, tested, and revised by the authors, who take full responsibility for the content of the software and of this paper. The framework architecture, the corpus schema, and the experimental design are the authors' own. Every bibliography entry has been verified against its DOI record.

Acknowledgements

RogueGPT is part of a broader research programme on AI-mediated misinformation conducted at Frankfurt University of Applied Sciences and IMT Atlantique. The companion tools CRED-1 (Loth, 2026a) and JudgeGPT, as well as the downstream provenance verification work Origin Lens (Loth, Conceicao Rosario, et al., 2026), are available as open-source software. The RogueGPT corpus is archived on Zenodo (Loth, 2026b). We thank the open-source communities behind Streamlit, PyMongo, and the Model Context Protocol SDK for the infrastructure that underpins this work.

References

Kreps, S., McCain, R. M., & Brundage, M. (2022). All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science*, *9*(1), 104–117. https://doi.org/10.1017/XPS.2020.37

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. https://doi.org/10.18653/v1/2022.acl-long.229

Loth, A. (2026a). *CRED-1: An open multi-signal domain credibility dataset* (Version v1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.18769460

Loth, A. (2026b). *RogueGPT stimulus corpus: A multilingual LLM-generated news dataset* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.18703137

Loth, A., Conceicao Rosario, D., Ebinger, P., Kappes, M., & Pahl, M.-O. (2026, May). Origin lens: Reclaiming trust on the AI-mediated web through on-device image provenance verification. *Companion Publication of the 2026 18th ACM Web Science Conference (WebSci Companion '26)*. https://doi.org/10.1145/3795513.3806658

Loth, A., Kappes, M., & Pahl, M.-O. (2026a). Eroding the truth-default: A causal analysis of human susceptibility to foundation model hallucinations and disinformation in the wild. *Companion Proceedings of the ACM Web Conference 2026 (WWW '26 Companion)*, 1125–1132. https://doi.org/10.1145/3774905.3795832

Loth, A., Kappes, M., & Pahl, M.-O. (2026b). Industrialized deception: The collateral effects of LLM-generated misinformation on digital ecosystems. *Companion Proceedings of the ACM Web Conference 2026 (WWW '26 Companion)*, 295–302. https://doi.org/10.1145/3774905.3795471

Loth, A., Kappes, M., & Pahl, M.-O. (2026c). The verification crisis: Expert perceptions of GenAI disinformation and the case for reproducible provenance. *Companion Proceedings of the ACM Web Conference 2026 (WWW '26 Companion)*, 980–988. https://doi.org/10.1145/3774905.3795484

Loth, A., Kappes, M., & Pahl, M.-O. (2024). *Blessing or curse? A survey on the impact of generative AI on fake news*. arXiv. https://doi.org/10.48550/arXiv.2404.03021

Macko, D., Moro, R., Uchendu, A., Lucas, J., Yamashita, M., Pikuliak, M., Srba, I., Le, T., Lee, D., Simko, J., & Bielikova, M. (2023). MULTITuDE: Large-scale multilingual machine-generated text detection benchmark. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, 9960–9987. https://doi.org/10.18653/v1/2023.emnlp-main.616

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, *8*(3), 171–188. https://doi.org/10.1089/big.2020.0062

Wang, W. Y. (2017). "Liar, liar pants on fire": A new benchmark dataset for fake news detection. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. https://doi.org/10.18653/v1/P17-2067

Wang, Y., Mansurov, J., Ivanov, P., Su, J., Shelmanov, A., Tsvigun, A., Whitehouse, C., Mohammed Afzal, O., Mahmoud, T., Sasaki, T., Arnold, T., Aji, A. F., Habash, N., Gurevych, I., & Nakov, P. (2024). M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2024)*, 1369–1407. https://doi.org/10.18653/v1/2024.eacl-long.83

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. *Advances in Neural Information Processing Systems*, *32*. https://arxiv.org/abs/1905.12616