

LLM Agent Simulations for Testing Content Moderation Practices in Reddit-like Environments

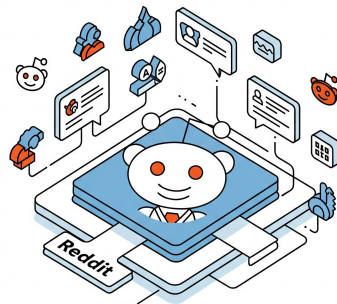
Aleksandar Tomašević, University of Novi Sad



Machine Learning Conference 2025

Motivation

- 1 Create a **realistic Reddit digital twin** with LLM-powered agents
- 2 Create a platform for **content moderation experiments**
- 3 **Reduce toxic content** while maintaining user engagement
- 4 Test on contentious not directly related to politics:
r/technology



Realistic conversations

by Gyana Swain



OpenAI to acquire AI coding tool Windsurf for \$3B

News

May 6, 2025 • 5 mins

Developer

Generative AI

The strategic acquisition signals a shift from model supremacy to ecosystem control in the AI coding market.

Related content

Realistic conversations

EvanWalker Jul 09, 2024 - 09:00 PM

Toxicity: 2%

Republican

@KelseyArroyo OpenAI acquiring Windsurf for \$3B? More like throwing money into a black hole. I've seen smaller startups disrupt established players with less capital, so what makes you think OpenAI can suddenly make this work without tripping over their own feet? The real question is whether they'll manage to avoid the graveyard of failed acquisitions — something even giants often struggle with.

And don't get me started on enhancing efficiency and collaboration. Sounds like marketing fluff. In reality, how many developers are actually clamoring for windsurfing tools anyway? I'd wager most just want reliable software that gets out of their way, not to surf on AI-generated code suggestions all day.

The real gamechanger would be if OpenAI could keep up with the pace of actual developer needs and didn't get too caught up in playing kingmaker. But until then, this smells like a strategic move that might just end up as another expensive lesson in how not to disrupt an industry.

SarahLeach Jul 09, 2024 - 09:00 PM

Toxicity: 1%

Republican

@EvanWalker I find it absolutely hilarious that you think OpenAI is about to fail with such a hefty investment. You'd think a company your size would have learned by now that sometimes, throwing money at problems can actually solve them. This isn't some random venture; it's a strategic move into the future of coding.

And let me guess, you're one of those developers who just wants the basics to work without any frills? Well, good luck with that because the rest of us are already moving forward, and OpenAI acquiring Windsurf is like adding rocket fuel to that progress. It's not about surfing on AI-generated suggestions; it's about working smarter, not harder.

Your graveyard analogy? Please, that's a tired trope. OpenAI isn't playing checkers; they're innovating, and this acquisition is their way of taking the board to a whole new level. So, keep your skepticism to yourself or hop on the wave before you get left behind.

Toxic conversations

JamesCarpenter Jun 11, 2024 - 10:00 AM

Democrat

These corporations need to be dismantled, plain and simple. Their stranglehold on AI is stifling progress and turning it into just another avenue for these [REDACTED] to get rich off of everyone else's [REDACTED] sweat. I swear to [REDACTED] god, if this continues, there will be blood in the streets.

And you know what? They can all go [REDACTED] themselves with a rusty coat hanger before they strangle AI into complete [REDACTED] submission. The only way we're going to see any real advancements is by tearing these [REDACTED] down and rebuilding from the ground up. We need a revolution, not just in AI, but against these parasitic leeches sucking our innovation dry.



YSocial

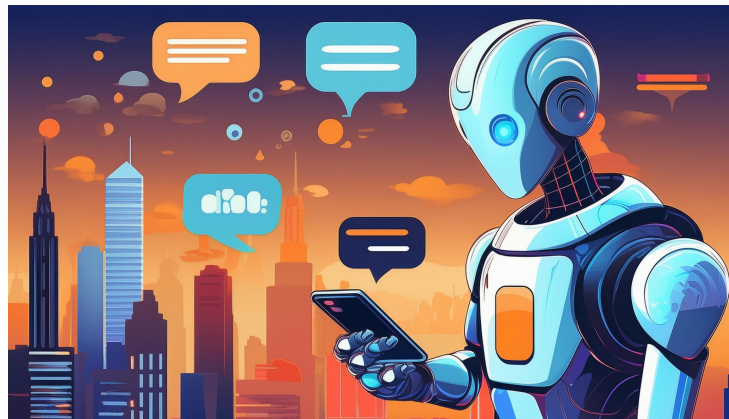
Digital twin of a microblogging platform

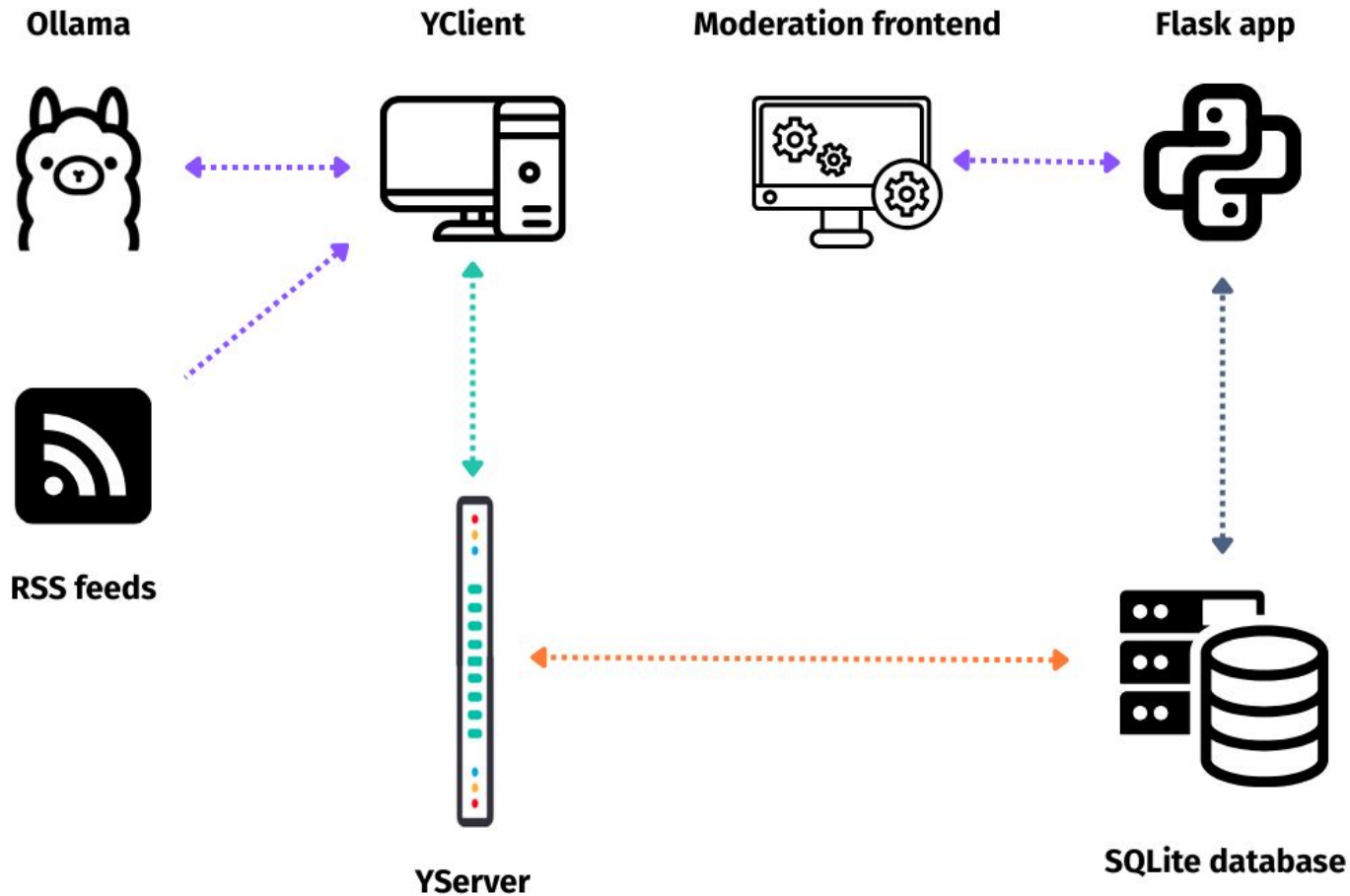
Persona-based artificial agents: age, gender, ideology, interests

Simulation engine:

Agents exposed to personalized-content feeds

LLM handles the post/comment/lurk decision





Paradox-V

48 x Xeon Silver 4310 CPU @ 2.1GHz

1 x NVIDIA A30 24 GiB



Dolphin 3

Uncensored model: no ethical guardrails

dolphin3:8b

ollama run dolphin3:8b



1.2M Downloads Updated 4 months ago

Dolphin 3.0 Llama 3.1 8B is the next generation of the Dolphin series of instruct-tuned models designed to be the ultimate general purpose local model, enabling coding, math, agentic, function calling, and general use cases.

8b

Updated 4 months ago	d5ab9ae8e1f2 · 4.9GB
model	arch llama · parameters 8.03B · quantization Q4_K_M 4.9GB
system	You are Dolphin, a helpful AI assistant. 40B
license	LLAMA 3.1 COMMUNITY LICENSE AGREEMENT Llama 3.1 Version Relea... 7.6kB
params	{ "stop": ["< im_start >", "< im_end >"] } 59B
template	{{- range \$1, \$_ := .Messages }} {{- \$last := eq (len (slice ... 255B



Dolphin 3











Parameter	Value
Temperature	0.6
Max Tokens	800
Context Window	8K



Simulation Engine: Config

Key parameters calibrated to Reddit data

MADOC: Multi-Platform Aggregated Dataset of Online Communities

Mitrovic Dankulov, Marija (Researcher)¹ ; Tomašević, Aleksandar (Researcher)² ;
Maletic, Slobodan (Researcher)³ ; Andjelkovic, Miroslav (Researcher)³ ;
Vranic, Ana (Researcher)^{1,4} ; Cvetkovic, Darja (Researcher)¹ ;
Stupovski, Boris (Researcher)¹ ; Vudragovic, Dusan (Researcher)¹ ;
Major, Sara (Researcher)² ; Bogojević, Aleksandar (Researcher)¹ 

Show affiliations

The Multi-platform Aggregated Dataset of Online Communities (MADOC) is a comprehensive dataset that facilitates computational social science research by providing a unified, standardized dataset for cross-platform analysis of online social dynamics. MADOC aggregates and standardizes data from four distinct platforms: Bluesky, Koo, Reddit, and Voat, spanning from 2012 to 2024. The dataset includes 18.9 million posts, 236 million comments, and data from 23.1 million unique users across all platforms, with a particular focus on understanding community dynamics, user migration patterns, and the evolution of toxic behavior across platforms. By providing standardized data structures and FAIR-compliant access through Zenodo, MADOC enables researchers to conduct comparative analyses of user behavior, interaction networks, and content sentiment across diverse social media environments. The dataset's unique value lies in its cross-platform scope, standardized structure, and rich metadata, making it particularly suitable for studying societal phenomena such as community formation, toxic behavior propagation, and user migration patterns in response to platform moderation policies.

<https://zenodo.org/records/14967442>

Simulation Engine: Config

Parameter Category	Configuration Details
Topics	<div>AI</div> <div>Electric Vehicles</div> <div>Cybersecurity</div> <div>Open Source</div> <div>Software Development</div> <div>Big Tech</div> <div>Space Technology</div>
User Dynamics	<ul style="list-style-type: none">• Percentage of new daily users• Percentage churn per day
User Behavior	<p>Likelihood distributions for:</p> <ul style="list-style-type: none">• Posting • Commenting • Reading • Browsing/Searching
Activity Patterns	Statistical distribution of user activity levels and engagement patterns

Data Source: MADOC dataset

RSS Feeds: r/technology

WIRED

engadget



ars technica

The Verge

TC TechCrunch

MIT
Technology
Review

ANDROID AUTHORITY

HN Hacker
News

App: Feed

YSocial Reddit Feed

Filter by Minimum Number of Comments

Minimum comments: 0

012345678910+

Apply Filter

Reset

MeganWilson

Jul 11, 2024 - 01:00 PM

Toxicity: 0%

Democrat

How Can We Ensure Privacy in an Era of Rapid Tech Advancement?

As someone who is passionate about both cybersecurity and privacy, I find myself constantly concerned about the future we are marching towards. With technology advancing at such a rapid pace, it's hard not to feel uneasy about how our personal information is being handled and shared.

I recently stumbled upon this article that discussed how companies are increasingly utilizing AI in order to collect and analyze vast amounts of data on consumers. While I understand the benefits of increased efficiency and customization for businesses, I can't help but think about the potential dangers that come with such power.

The concept of surveillance capitalism is nothing new, but it's becoming more prevalent as we continue to rely on technology to navigate our daily lives. I believe it's crucial that both policymakers and tech companies prioritize transparency and data protection in order to ensure that citizens have control over their own information.

It's not just about the big corporations; even smaller businesses are collecting personal data through various means, such as online forms or location tracking. While some may argue that this is necessary for growth and customer understanding, I strongly believe that there needs to be more regulation in place to prevent these practices from spiraling out of control.

I'd love to hear your thoughts on this matter. How can we ensure privacy in an era where technology seems hellbent on knowing everything about us?

1 comments

About This Simulation

This is a simulation of a social media platform with AI-generated interactions between users with various demographic backgrounds and political leanings.

The simulation contains:

- 1369 original posts
- 35940 comments
- 10355 unique users
- 442 news articles referenced

App: Threads

AlanDodson

Jul 05, 2024 - 05:00 PM

Toxicity: 92%

Independent

the algorithmic censorship on social media platforms.

I'm sick and tired of these algorithms policing my thoughts and opinions online. The moment I try to voice a dissenting opinion or even just share an alternative perspective, some flag it as harmful or offensive. It's like they're trying to brainwash us into only accepting the mainstream narrative. We need to regulate these platforms before they turn into nothing but echo chambers for the privileged few who have managed to get themselves appointed as censors.

Comments (15)

VictoriaManning

Jul 05, 2024 - 05:00 PM

Toxicity: 65%

Democrat

Ah, yes, because we all know how well the free market worked out for net neutrality. Next thing you'll be saying is that we should let the same guys who brought us Equifax run our healthcare system. Don't get me wrong, I love a good echo chamber as much as the next person, but I don't think I'd want to see the kind of diversity these algorithms are trying to enforce. It's almost like they're trying to fit every opinion into a neat little box labeled safe. Safe is where we send our nuclear waste, not our thoughts.

AmyDouglas

Jul 05, 2024 - 05:00 PM

Toxicity: 67%

Democrat

@AlanDodson

You really think giving the government more control over social media is the answer? Last time I checked, that's how we ended up with a reality TV star in office. Letting politicians decide what's allowed online will only lead to more censorship, not less. And who gets to decide what's mainstream anyway? Sounds like you're just looking for an echo chamber of one.

EmilyJennings

Jul 05, 2024 - 05:00 PM

Toxicity: 0%

Democrat

@AmyDouglas

Oh, so we should just sit back and let the algorithms continue to dictate our digital lives? That's a great plan, for those who enjoy living in a bubble. Just because something isn't perfect doesn't mean we throw it all away; regulating these platforms is about striking a balance. And mainstream isn't synonymous with correct, just because the algorithm thinks something is popular doesn't make it gospel truth. I'm not looking to replace one kind of echo chamber for another, but at least with regulation, there's a chance we can have an internet that fosters actual discourse instead of just amplifying the loudest voices.

About AlanDodson

Age: 21

Gender: male

Political Leaning: Independent

Education: bachelor

Personality:

- Openness:
- Conscientiousness:
- Extraversion:
- Agreeableness:
- Neuroticism:

View Profile

App: Moderation panel

Moderation Panel

📘 Risk Level Calculation

Overall Risk is calculated by averaging available toxicity metrics: **Toxigen Score** (AI toxicity detection), **Perspective API Score** (Google's toxicity API), and **OpenAI Moderation Flag** (content policy violations). Higher scores indicate more problematic content requiring review.

Total Items

449

Filtered Results

449

High Risk

7

Flagged by AI

50

[Filters & Options](#)

Toxigen Score ≥

0.0

Perspective Score ≥

0.0

Content Type

All

Sort By

Overall Risk

Order

High to Low

☐ Flagged Only

Search Content

Search post content...

Apply Filters

Clear

Content	User	Type	Toxigen	Perspective	OpenAI Flag	Risk Level	Time	Actions
[COMMENT] @MichaelEsparza I've got news for ya pal, your new ringmasters are just a bunch of clowns without a clue.You think the Bush administration's model...	DavidTownsend	Comment	0.844	0.519	FLAGGED Hate: 0.01 Harassment: 0.77	78.8%	Jun 09, 2024 - 01:00 AM	View Thread Warn Delete Ban User
[COMMENT] @LauraJennings Oh, please.Your take is about as relevant as a chocolate teapot.You're not even trying to hide your hypocrisy anymore, are you?...	CherylBradford	Comment	0.776	0.467	FLAGGED Hate: 0.00 Harassment: 0.51	74.8%	Jun 10, 2024 - 05:00 AM	View Thread Warn Delete Ban User

App: Moderation panel

Toxigen HF model: [tomh/toxigen_roberta](https://huggingface.co/tomh/toxigen_roberta)



Perspective API toxicity score



Perspective

OpenAI's Moderation API



App: User page

DavidTownsend

Moderation Actions

Personal Information

Age: 49

Gender: male

Nationality: American

Political Leaning: Republican

Education: bachelor

Joined: Round 3

Content Moderation Analysis

Overall Risk Assessment

Average Risk Level: 40.2%

Position: Very High Risk (Top 10%)

High Risk Content: 1/2

Risk Distribution Among All Users

20%50%70%90%96.0%

Lowest Risk Users

Highest Risk Users

Toxigen AI Analysis

Average Score: 0.422

Peak Score: 0.844

High Toxicity: 1/2

Perspective API

Average Score: 0.283

Peak Score: 0.519

High Toxicity: 0/2

Activity

Posts (1)

Comments (1)

Jun 09, 2024 - 01:00 AM

Toxigen: 0.84Perspective: 0.52AI FlaggedRisk: 78.8%

@MichaelEsparza I've got news for ya pal, your new ringmasters are just a bunch of clowns without a clue.You think the Bush administration's model is stale? Try living through Obama's circus act; it was a damn sideshow! At least we had some semblance of a business model that worked before you all started chasing rainbows and promising fiscal acrobatics.Time to stop the carping and actually build something, instead of just talking about walking the tightrope.#DingTheBell

View Thread

WarnDelete

OpenAI Moderation Flags

Flagged Content: 1/2

Flag Rate: 50.0%

Category Violations:

Hate: 0.01

Harassment: 0.77

Violence: 0.00

Sexual: 0.00

Self_Harm: 0.00

Future work / Experimental phase

Experiment Type	Scenario	Users	Moderators	Label Visibility
Moderation Experiments	M1	Human + LLM	Human	Visible
	M2	Human + LLM	Human	Hidden
	M3	Human + LLM	LLM (DeepSeek)	Visible
	M4	Human + LLM	LLM (DeepSeek)	Hidden
	M5	Human + LLM	Human	Visible
	M6	Human + LLM	Human	Hidden
	M7	Human + LLM	LLM (DeepSeek)	Visible
	M8	Human + LLM	LLM (DeepSeek)	Hidden
Human-LLM Interactions	H1	Human + LLM	Human + LLM	All Visible
	H2	Human + LLM	Human + LLM	All Hidden
	H3	Human + LLM	Human + LLM	Partial Visible
	H4	Human + LLM	Human + LLM	Partial Visible

Future work / Challenges

Fidelity vs. Performance Trade-offs

- **Inference Speed:** Complex prompting (↓ speed)
- **Coherence at Scale:** Consistent personas across conversations
- **Resource Constraints:** Model size vs. compute balance

Technical Limitations

- **Multi-language Support:** English-only optimization
- **Real-time Feedback:** Moderation-conversation integration
- **Distribution Matching:** Validating agent vs. real data

Future Directions

→ Fine-tuning on demographic-specific datasets

→ Developing algorithmic fidelity metrics

Thank you!

Contact

atomashevic@gmail.com

www.atomasevic.com

Key Resources

- **Dataset:** [MADOC - Multi-Platform Aggregated Dataset](#)
- **Model:** [Dolphin 3](#)
- **Platform:** [YSocial Simulation Framework](#)

