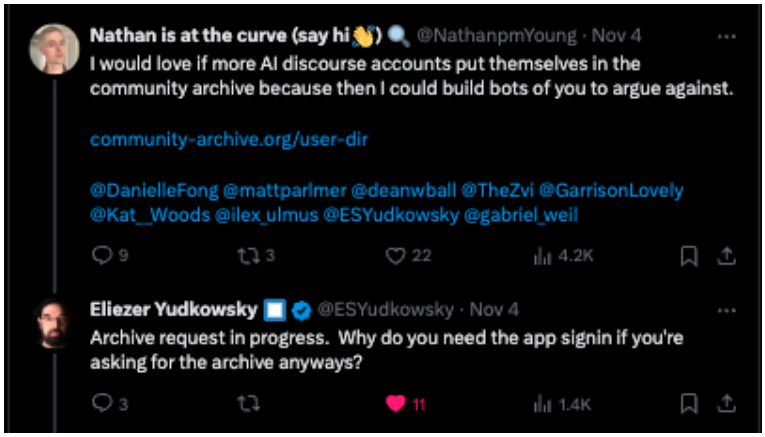
Project Proposal: Twitter/X Data to Naptha Personas - Building Infrastructure through TPOT

What are you trying to do?
Create a general-purpose pipeline that can transform any Twitter/X data export into deployable Naptha personas, while bootstrapping adoption by processing Community Archive's 12.4M tweet dataset from the “TPOT” community on X, which many in and around machine learning call home. This initial dataset represents a uniquely valuable nexus of machine learning thought leadership, including influential figures like @repligate (creator of Loom LLM interfaces), Emmett Shear former CEO of OpenAI, and Zvi Moskowitz, influential blogger on AI Safety. There is significant demand to turn these data into agents one can interact with:



The project involves seeks to turn that demand into engagement with Naptha.ai:
- Converting tweet threads and conversations into ChatML format suitable for LLM training
- Extracting preference signals from like patterns and timestamps
- Hydrating the dataset with linked media and external content
- Creating universal tooling for deploying Twitter/X content as Naptha personas

How is it done today, and what are the limitations?
Currently, Twitter/X provides data exports in JSON format, but this raw data isn't directly suitable for:
- Training conversational AI models (requires thread reorganization and ChatML formatting)
- Preference learning (requires structured extraction of engagement signals)
- Document-grounded chat (requires hydration of external links and media)
- Direct deployment as personas on agent platforms

These limitations affect both individual users wanting to create personas from their content and communities like TPOT seeking to preserve and extend their collective knowledge. The current friction prevents widespread adoption of AI personas based on Twitter/X content.

What is new in your approach and why do you think it will be successful?
Our approach combines:
1. A universal pipeline that can process any Twitter/X data export into AI-ready formats
2. Deep integration with Naptha's hub through automated persona deployment tooling
3. Strategic initial deployment with TPOT community data to demonstrate value
4. Collaboration with key community members (e.g., Xiq, creator of community archive)

This will succeed because:
- The tooling will be generally applicable to any Twitter/X data export
- We're proving the concept with a high-value initial dataset (TPOT community)
- Twitter/X agents are gaining significant attention (e.g., Truth Terminal's recent a16z funding)
- The pipeline will enable anyone to create and deploy personas from their Twitter/X presence
- We have existing relationships with key community members to ensure effective initial deployment
- The “TPOT” community (in which many ML engineers participate) is highly invested in self-knowledge and self-actualization through technology inclusive of machine learning, which we expect to drive organic adoption as TPOT-based Personas are deployed into applications.

Who cares? If you succeed, what difference will it make?
Primary beneficiaries:
1. Naptha:
   - Gains a general-purpose tool for turning Twitter/X content into personas
   - Initial population of their hub with high-value ML community personas
   - Advances their vision of decentralized AI infrastructure through accessible tooling
   - Positions their platform as the go-to solution for Twitter/X persona deployment
2. Twitter/X users broadly:
   - Can easily convert their content into interactive personas
   - Gain tools for preserving and extending their online presence
3. TPOT community:
   - Serves as initial proof of concept
   - Creates a new platform for ML discourse and experimentation
4. AI developers:
   - Access to tools for processing Twitter/X data into training formats
   - Datasets of pre-existing twitter data in ChatML format published to Huggingface
   - Easy deployment of personas from any Twitter/X dataset
   - Foundation for building larger agent ecosystems

What are the risks and payoffs?
Risks:
- Platform integration challenges requiring iteration on design
- Potential perception concerns around AI personas
- Maintaining pipeline reliability as Twitter/X API evolves

Payoffs:
- Universal infrastructure for Twitter/X data processing
- Driving adoption of Naptha's platform through both ML community and general Twitter/X users
- Establishing foundation for ongoing agent ecosystem development
- Positioning Naptha's hub as a key platform for social media presence extension
- Creating reusable tools for the broader AI community

How much will it cost?
$12,000 total budget:
- 400 hours combined work at $30/hour
- Split between two developers over approximately 5 weeks
- Expertise combines ML engineering and theoretical mathematics

How long will it take?
5 weeks, with four major milestones:
1. Detailed implementation plan and design document (Week 1)
2. ChatML formatting pipeline and initial model fine-tuning (Week 2-3)
3. Dataset hydration with external content (Week 3-4)
4. Preference dataset creation and deployment tooling (Week 4-5)

What are the mid-term and final "exams" to check for success?
Technical Metrics:
- Complete processing of 12.4M tweet TPOT dataset and publication on Huggingface
- Successful deployment of example personas on Naptha
- Pipeline maintaining updates as Community Archive grows
- Demonstrated ability to process new Twitter/X data exports
- Performance metrics for deployed personas

Engagement Metrics:
- TPOT community interaction with deployed personas
- Number of new personas deployed using our tooling
- Derivative datasets and models on platforms like Huggingface
- Adoption by Twitter/X users outside TPOT community

Key Deliverables:
1. Design document detailing implementation plan and Naptha integration strategy
2. ChatML-formatted dataset and fine-tuned model
3. Hydrated dataset with external content
4. Preference dataset and deployment tooling
5. Example personas deployed on Naptha platform
6. Documentation and maintenance scripts
7. General-purpose pipeline for processing any Twitter/X data export

About the team:
George Walker is a technologist, software developer, pre-training dataset influencer, and long-time community organizer in TPOT, for which he co-founded VibeCamp, started a tpot community in Portugal, and has thrown many parties for some of the biggest names in ML.

Peter Graziano is a mathematician and data scientist specializing in machine learning applications, with a strong foundation in partial differential equations, numerical analysis, scholastic theology, classical languages, and data munging in Python.