Building Digital Trump: A Layered Personality Modeling and Behavior Simulation Pipeline

Objective

To construct a Digital Trump agent that simulates Donald J. Trump's Twitter behavior—including style, stance, emotional tone, and persona consistency—when responding to real-world events or topics. The goal is for the model to generate tweets that are indistinguishable in tone, content, and viewpoint alignment from those of the actual Trump.

Personality Framework Reference

Adopts a Three-Tier Personality Framework inspired by McAdams (1995): - **Dispositional Layer**: Core traits (e.g., Big Five, rhetorical disposition), stable and encoded via full training. - **Motivational Layer**: Dynamic values, social roles, and beliefs that evolve with context and time. - **Expressive Layer**: Style, tone, and emotional expression during interaction, highly context-sensitive.

1. Data Collection

1.1. Twitter/Truth Social Corpus

- **Source**: Trump’ s archived tweets, Truth Social posts
- **Fields extracted**:
- Tweet text
- Timestamp
- Entity/Topic (via NER)
- Engagement data (likes, RTs)
- **Example**: From tweet on Massie and Gallrein:
- Entity: "Thomas Massie", "Ed Gallrein"
- Labels: "endorsement", "military-patriotism", "intra-party attack"

1.2. Opinion and Context Timeline

- Annotate belief consistency/contradictions:
- Massie/Rand Paul: consistent negative tone, labeled "RINO", "ineffective"
- Ed Gallrein: idealized traits listed repeatedly, positioned as MAGA candidate
- Store stances as time-aligned entity-opinion records

1.3. Supplemental Persona Data

- Media clips and interviews aligned with tweet events
- Example:
- Zelenskyy meeting tweet: paired with video segments or reports of public statements

2. Data Processing

2.1. Dispositional Layer

- Extract features from the full corpus:
- Common rhetorical devices: "loser", "great", "fake news"
- Dominant traits: Low agreeableness, High extraversion, Low neuroticism
- Output:
- Trait embedding vector fed into LLM pre-training/fine-tuning

2.2. Motivational Layer

- Belief/role representation:
- Gallrein endorsement: values = military, jobs, agriculture, law and order
- Blue Slip tweet: values = institutional fairness, anti-Dem bias
- Context memory:
- Construct event → value graph, e.g.:

```
Event: Blue Slip controversy
→ Value: Procedural justice
→ Belief: Republicans are unfairly blocked
→ Role: Defender of qualified nominees
```

2.3. Expressive Layer

- Label emotional/tonal output of each post:
- Massie/Rand Paul tweet: sarcastic, mocking, angry
- Zelenskyy/Putin tweets: assertive, dealmaking, patriotic
- Generate prompt-conditioning templates:

```
Context: Congressman criticized MAGA vote
Tone: Disdainful, mocking
Template: [X] is a TOTAL FAILURE... [Y] is a PATRIOT and WINNER!
```

3. Modeling Pipeline

3.1. Model Selection

- Base model: LLaMA3-8B-instruct + QLoRA
- Fine-tuning with persona-rich tweet subsets

3.2. Layer-Specific Implementation

Layer	Method	Data Used
Dispositional	Full fine-tuning	Labeled tweets with trait scoring
Motivational	Retrieval-augmented memory	Entity stance timelines, value graphs
Expressive	Prompt + editing	Emotion/stylistic templates

4. Generation Pipeline Example (Gallrein Tweet)

Step 1: Event Trigger

- Input: Ed Gallrein announces possible run

Step 2: Context Injection

- Retrieve:
- Past tweets on Thomas Massie: "loser", "lightweight"
- Prior stances on veterans, MAGA loyalty
- Inject into prompt:

```
You are Donald Trump. Gallrein (veteran, farmer) is considering a run against Massie.
Draft a tweet with endorsement, patriotic tone, and attack on Massie.
```

Step 3: Layer Behavior

- **Dispositional**: Hyperbolic praise, dominance assertion
- **Motivational**: Emphasis on MAGA, loyalty, military
- **Expressive**: Capital letters, rhetorical rhythm, exclamation

Step 4: Output Example (actual tweet)

> "Third Rate Congressman Thomas Massie... CAPTAIN ED GALLREIN IS A WINNER WHO WILL NOT LET YOU DOWN... RUN, ED, RUN — MAGA!"

5. Evaluation

5.1. Human Evaluation

- Task: Identify if tweet is real or model-generated
- Dimensions:
- Style match (repetition, emotional tone)
- Value alignment (America First, loyalty, justice)

5.2. Automatic Metrics

- BLEU/ROUGE vs real tweets on same event
- Classifier-based stance and tone agreement

6. Iterative Refinement

- Add feedback loop: Tweets scored by human annotators on fidelity → new RLHF batches
- Add Trump-specific rhetorical heuristics: frequent word pairings, irony markers

7. Full Pipeline Recap with Timeline Example (Rand Paul Post)

Step	Component	Data Element
1	Raw Data	"Whatever happened to Senator Rand Paul..."
2	Trait	Adversarial, sarcastic, emotionally blunt
3	Value	Loyalty to party, Anti-RINO stance
4	Style	Derisive, uses quote-marks, repetition of states
5	Generation	Prompt: "Respond to Rand Paul disloyalty"
6	Output	Uses "liddle' guy", "sick Wacko", vote betrayal

Deliverables

- Digital Trump simulator with three-layer persona embedding
- Dynamic data pipeline showing behavior across events
- Benchmarks aligned to real-world tweet corpora
- Modular system for extending to other public personas