Proposal for an LLM-Driven Simulation of Misinformation Spread

1 Introduction

As information warfare and misinformation campaigns become increasingly utilized in national security, evolving into socio-technical campaigns that leverage networks to influence public opinion, accurately simulating and modeling these dynamics at scale has become imperative for defensive strategies. In this proposal, we delineate a detailed design for an agent-based social framework that integrates large-language models (LLMs) with empirical psychological decision rules, enriched by real-world human data, and validated through well-defined ground truths. The framework will allow exploration of rumor propagation, counter-message effectiveness, and emergent social dynamics.

2 Agent Design

2.1 Persona Profiles

The agent will be instantiated with a multi-dimensional profile drawn from empirical surveys and collected demographics from real-world populations, encapsulating the following:

- **Demographic attributes** of individuals, including age, education level, residential context(urban/rural), income level, etc.
- **Psychometric factors** that encompass political orientations, emotional baseline (valence/arousal metrics), cognitive bias (including confirmation bias and authority bias), social conformity threshold, etc.

2.2 Dual-Phase Decision Model

We plan for the agent's behavior to derive from a two-stage process:

1. **Quantitative Evaluation**:
    - **Belief Score**: a weighted sum of source credibility, ideological alignment, emotional framing intensity, prior exposures, and confirmation bias.
    - **Share Propensity**: incorporates the belief score, peer-influence factors (such as neighbor adoption), authority bias, and emotional valence.
2. **LLM-Driven Text Generation**
    - If an agent exceeds a pre-determined threshold, text generation will be triggered for the LLM to generate a contextually coherent and persona-consistent message text.

By separating decision rules from text generation, the system minimizes high-cost LLM invocations, performing arithmetic and Boolean checks for most agents with each step, and only triggers LLMs for agents whose share propensity exceeds the pre-determined threshold. The dual-phase architecture thus ensures both scalability and persistence of linguistic realism in a shared context.

2.3 Memory Dynamics and Cognitive Biases

Agents are expected to maintain finite memory buffers that track recent exposures and debunk events, enabling functions such as:

1. **Decay functions** to model forgetting and permit re-forgetting.
2. **Reinforcement bias**, where repeated alignment with prior beliefs amplifies the susceptibility of the agent to similar narratives.

3. **Debunking diminution**, when an agent encounters a corrective message about a particular claim, it decreases its belief score for that claim in subsequent evaluations.

2.4 Social Influence and Conformity

Agents will continuously track the fraction of neighbors who either believed or shared a message. If this exceeds a defined **conformity threshold**, they will receive a boost to their **share score**, reflecting the individual's proneness to the bandwagon effect and echo chamber reinforcement.

2.5 Emotional Framing and Arousal Modulation

Messages will be tagged with **affective frames** (e.g., "fear", "anger", "excitement") quantified through arousal-valence metrics. High-arousal framed messages will amplify both the **belief score** and **share propensity** via predefined weights.

2.6 LLM-Driven Text Generation

When an agent decides to share a message, an LLM crafts a personalized post, such as one shown below:

```
prompt = f"""
You are a {node['age']}-year-old {node['education']} graduate with
political={node['political']:.2f}.
You BELIEVE the claim: \"{message.content}\".
Write 1–2 sentences to post, reflecting your style.
"""
post_text = llm_api.call(prompt)
```

3 Real-World Data Integration

In order to ground the agents in the simulation and mirror realistic behavior patterns in the real world, we will seed model components with empirical data as below:

Data Type	**Sources**
Demographic Distributions	[National census microdata (e.g. IPUMS), local statistical bureaus]
Media Trust & Credibility	[Pew Research Center's media trust surveys; European Social Survey questions on misinformation susceptibility]
Social Network Topologies	[SNAP datasets: Twitter, Reddit comment graphs]
Sharing Behavior Dynamics	[Twitter retweet cascades (via Academic API); Facebook sharing patterns (anonymized)]
Psychometric Profiles	[myPersonality dataset (Big Five scores); World Values Survey; Moral Foundations Survey]
Message Framing Effects	[Experimental psychology papers on fear appeals and persuasion (valence/arousal norms)]

4 Ground Truth Definition

4.1 Truth Labels

- **Binary truth labels** for each message based on fact-checker databases (e.g., Snopes, PolitiFact).
- **Continuous confidence** for partial truths that carry confidence scores.

4.2 Benchmark Spread Curves

- Historical propagation time series (e.g., retweet time series from Twitter)serving as empirical benchmarks for **infection curves**.

4.3 Behavioral Outcome Metrics

- **Belief adoption rates** will be compared against previous survey experiments to measure false rumor acceptance.
- **Reshare propensities** matched to social platform statistics for the sharing of debunked posts.

4.4 Calibration Targets

- Use the **R-effective** to calculate the average number of secondary belief adoptions per originator, and calibrate this against real-world metrics (e.g., average retweet counts per user, the spread of a hashtag).
- Compare known event timelines (e.g., misinformation spread around large-scale crises) via **peak reach timing**.

5 Simulation Architecture

1. **Data Ingestion Layer**: ETL pipelines to load census, survey, network, and fact-check data.
2. **Agent Initialization**: sample personas using integrated datasets.
3. **Network Assembly**: graph construction from SNAP or parameterized synthetic small-world/scale-free generators.
4. **Simulation Engine**: orchestrates asynchronous event loops, rule-based score computations, and batched LLM invocations.
5. **State Management**: Remote Dictionary Server for inbox message handling and PostgreSQL for long-term metric storage.

6. **Scalability Layer**: distribute LLM API calls and agent updates across compute clusters.
7. **Visualization**: real-time dashboards (e.g., Streamlit, Dash) to monitor diagnostics and inspect individual agent logs.

6 Metrics and Validation

We will track and compare the following against the ground truth:

1. **Infection curves** vs. historical cascades by RMSE
2. **Belief-share conversion rates** vs. survey data
3. **Community-level spread** vs. empirical subgraph statistics
4. **Echo chamber polarization**: divergence of opinions over time within communities

We plan to use optimization methods (e.g., grid search, Bayesian optimization) to tune psychological weights and thresholds to minimize discrepancies.

7 Conclusion

This framework plans to bridge the gap between social-science modeling with the flexibility of LLMs, anchored in real-world datasets and validated against clear and easily obtainable ground truth. By iterating on data integration, rule calibration, and ethical oversight, we plan to build a robust simulation platform to inform public communication planning.