Proposal outline:

1. Organize the data in a SQL database equipped with an on-device NL2SQL agent, allowing me to perform rapid exploratory analysis by just asking questions in plain English (e.g. what are the top 100 cities with tweets mentioning guns?) The tool will streamline gaining statistical insight from the dataset.
2. Embed the tweets using a sentence transformer model (e.g. ModernBERT). This would enable a range of applications including outlier identification, semantic clustering, and retrieval-augmented generation. By creating embeddings for each tweet, I hope to facilitate quick semantic searches and topic-level groupings within the dataset.
3. Performing hierarchical topic modeling on the dataset by utilizing embeddings, followed by the use of an on-device LLM for generating and refining cluster names, followed by an informative data visualization.
4. Producing a cleaner and richer version of the dataset by removing unrelated data points and adding topic tags, X-24-US-Election 2.0.

I'm well aware that this work is not proposing a new research question but would be quite necessary for any reliable subsequent work. If time would allow, I would be interested in creating an agentic model for detecting cognitive biases and logical fallacies in threaded conversations. This work could potentially help with identifying (and possibly mediating) polarizing or misinformed conversations.