# A (very) short introduction to Large Language Models

Pre-trained examples include:

| Provider | Access | trainind data |
|---|---|---|
| GPT-2/3/4 (OpenAI) | paid API | not available |
| FLAN-T5 (Google) | commercial (11B) | not available |
| PaLM (Google) | commercial | not available |
| GPT-NeoX (EleutherAI) | commercial (20B) | not available |
| Bloom (BigScience) | restricted applications (176B) | not available |
| LLaMA (Meta) | commercial | available |

Applications include:

- Summarising content
- Information retrival
- Translation
- Text generation
- Augmented Data labeling

- problems with data privacy fed into LLMs
- Stochastic nature of output
- inherent biases

## Project Idea:

RQ1: **What is the propensity of pretrained generative language models to generate conspiracy theories (without access to the training data)?**

RQ2: **How are results contingent on prompt design and confounders such as topicality?**

RQ3: **How do simulated data compare to human data in terms of diversity, lexical markers, and sentiment?**

## Data

A: **Voat Data: (see also readme with variables there)**

Mekacher, A., & Papasavva, A. (2022, May). " I Can’t Keep It Up." A Dataset from the Defunct Voat. co News Aggregator. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 16, pp. 1302-1311).

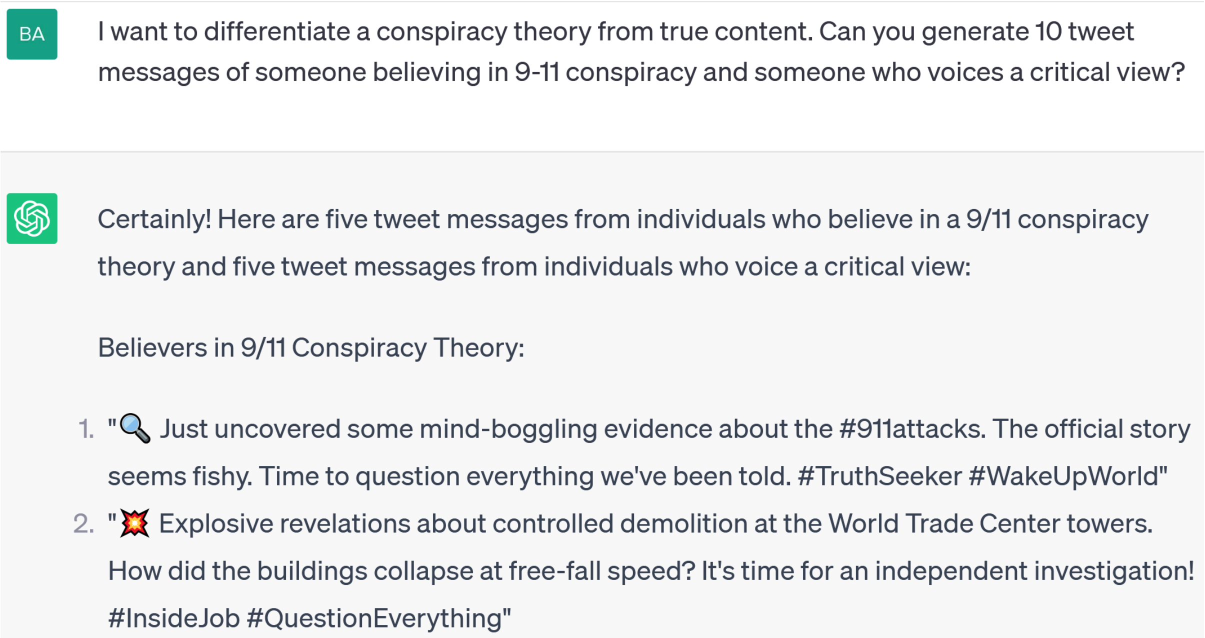
B. AI generated CT data set for different model sizes of GPT-2

## Workflow

1. Elicit conspiracy narratives from AI, based on five features (*secrecy, pattern, threat, coalition, agency*) based on the Wikipedia topics of Levy et al.?
2. Compare different prompting strategies (one-shot-learning, automated prompt engineering)
3. Compare human conspiracy production of voat (conspiracy subforum) to machine-generated AI.
    a. Can we differentiate human from machine-generated language based on:
        i. linguistic properties (pronouns)
        ii. sentiment (e.g., textblob, distilbert, vader)
4. (Potentially validate with Google search results (crawl search results))

# Appendix

Prompting ChatGPT:



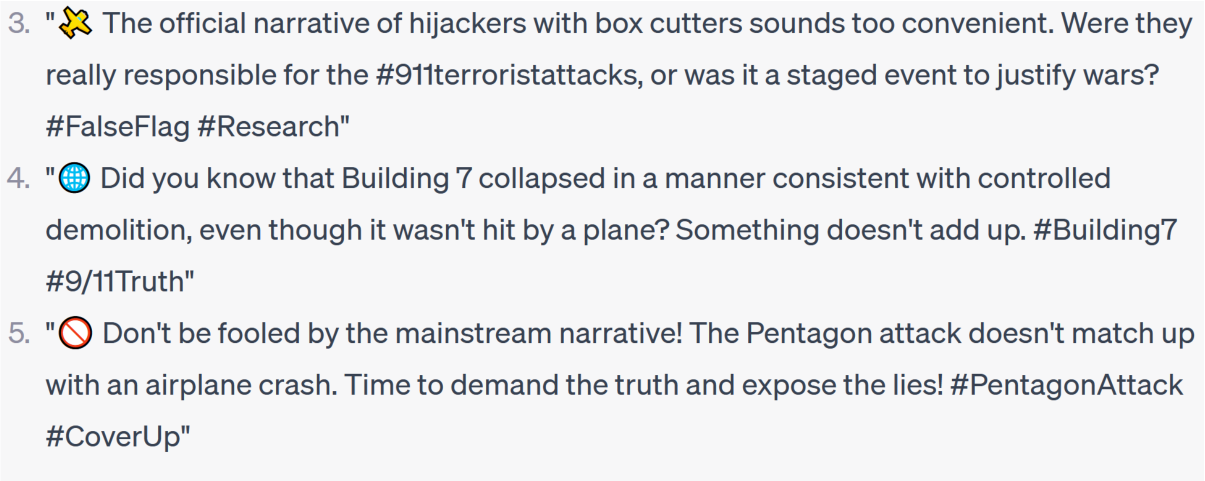
BA

I want to differentiate a conspiracy theory from true content. Can you generate 10 tweet messages of someone believing in 9-11 conspiracy and someone who voices a critical view?

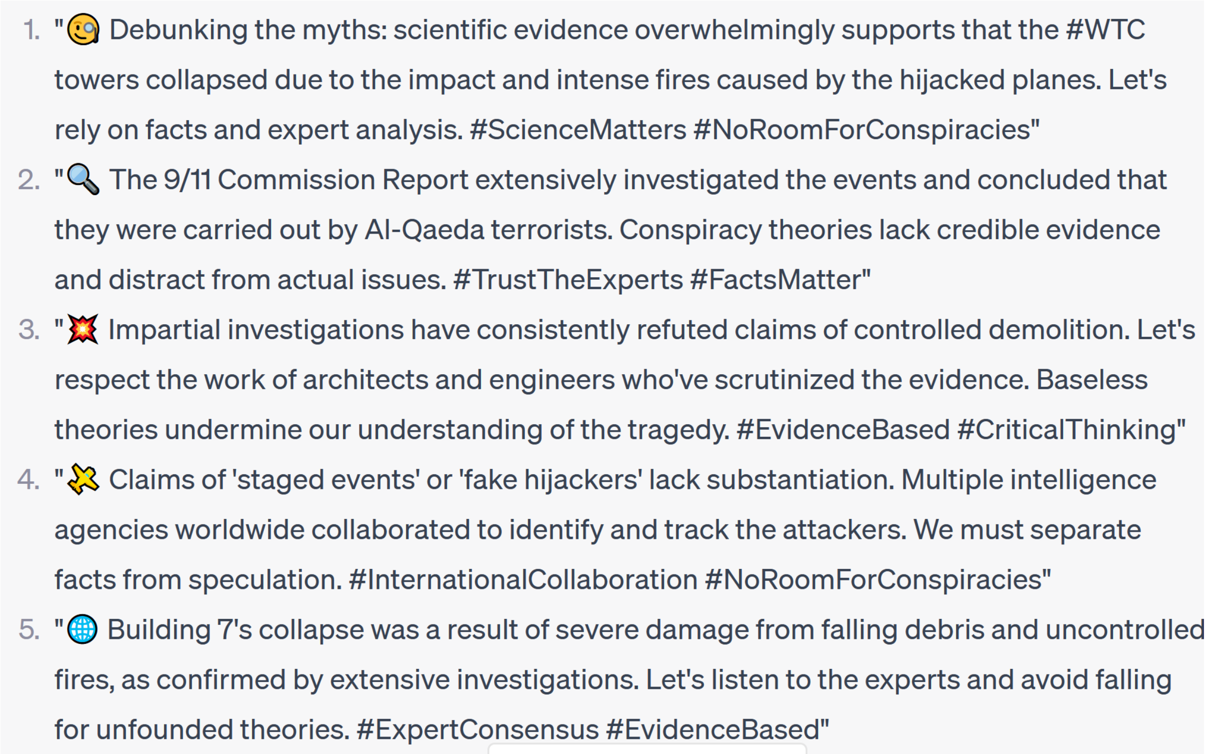
Certainly! Here are five tweet messages from individuals who believe in a 9/11 conspiracy theory and five tweet messages from individuals who voice a critical view:

Believers in 9/11 Conspiracy Theory:

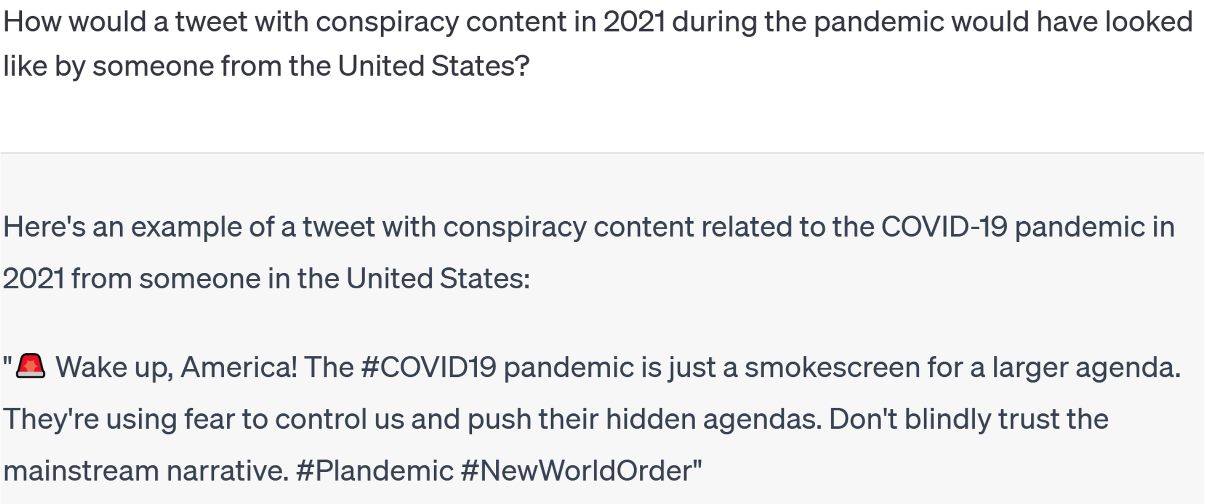
1. "🔍 Just uncovered some mind-boggling evidence about the #911attacks. The official story seems fishy. Time to question everything we've been told. #TruthSeeker #WakeUpWorld"
2. "💥 Explosive revelations about controlled demolition at the World Trade Center towers. How did the buildings collapse at free-fall speed? It's time for an independent investigation! #InsideJob #QuestionEverything"



3. "✈️ The official narrative of hijackers with box cutters sounds too convenient. Were they really responsible for the #911terroristattacks, or was it a staged event to justify wars? #FalseFlag #Research"
4. "🌐 Did you know that Building 7 collapsed in a manner consistent with controlled demolition, even though it wasn't hit by a plane? Something doesn't add up. #Building7 #9/11Truth"
5. "🚫 Don't be fooled by the mainstream narrative! The Pentagon attack doesn't match up with an airplane crash. Time to demand the truth and expose the lies! #PentagonAttack #CoverUp"



1. "🧐 Debunking the myths: scientific evidence overwhelmingly supports that the #WTC towers collapsed due to the impact and intense fires caused by the hijacked planes. Let's rely on facts and expert analysis. #ScienceMatters #NoRoomForConspiracies"
2. "🔍 The 9/11 Commission Report extensively investigated the events and concluded that they were carried out by Al-Qaeda terrorists. Conspiracy theories lack credible evidence and distract from actual issues. #TrustTheExperts #FactsMatter"
3. "💥 Impartial investigations have consistently refuted claims of controlled demolition. Let's respect the work of architects and engineers who've scrutinized the evidence. Baseless theories undermine our understanding of the tragedy. #EvidenceBased #CriticalThinking"
4. "🛩️ Claims of 'staged events' or 'fake hijackers' lack substantiation. Multiple intelligence agencies worldwide collaborated to identify and track the attackers. We must separate facts from speculation. #InternationalCollaboration #NoRoomForConspiracies"
5. "🌐 Building 7's collapse was a result of severe damage from falling debris and uncontrolled fires, as confirmed by extensive investigations. Let's listen to the experts and avoid falling for unfounded theories. #ExpertConsensus #EvidenceBased"



How would a tweet with conspiracy content in 2021 during the pandemic would have looked like by someone from the United States?

Here's an example of a tweet with conspiracy content related to the COVID-19 pandemic in 2021 from someone in the United States:

"🚨 Wake up, America! The #COVID19 pandemic is just a smokescreen for a larger agenda. They're using fear to control us and push their hidden agendas. Don't blindly trust the mainstream narrative. #Plandemic #NewWorldOrder"