Prompt Compass

A Methodological Approach to Evaluating the Use of LLMs in SSH research

Dr. Erik Borra
borra@uva.nl
System prompt: You are an advanced classifying AI. You are tasked with classifying the sentiment of a text. Sentiment can be either positive, negative or neutral.

Prompt: Classify the following social media comment into either 'negative', 'neutral' or 'positive'. Your answer MUST be either one of ['negative', 'neutral', 'positive']. Your answer must be lowercase.

Text: {user_input}

Answer:
Extract the important entities mentioned in the text below. First extract all country names, then extract all organizations, then extract all people names, then extract specific topics which fit the content and finally extract general overarching themes

Desired format:
Country names: <comma_separated_list_of_country_names>
Organization names: <comma_separated_list_of_organization_names>
People names: -||-
Specific topics: -||-
General themes: -||-

Text: {text}
Country names: United States, Germany


People names: Fred Havemeyer, Ryan Abbott, Jane Ginsburg

Specific topics: Copyright Law, Artificial Intelligence (A.I.), Intellectual Property, ChatGPT Bing Chat, Data Licensing Agreements, Fair Use Doctrine, Generative A.I. Technologies, Venture Capital, Legal Landscape in A.I.


Hypothesis: An LLM can find narratives in multiple posts with few-shot learning.

Prompt:

You're an expert in narratology. Narrative is a series of claims that make up a story that serves a specific purpose. Below is an example of a narrative:

```
{
  "Title": "The West controls Ukraine and uses it to its advantage",
  "Characters": {
    "West": "Potentially referring to Western countries or alliances like NATO",
    "Ukraine": "The nation caught in the implied manipulation or control"
  },
  "Plot": "A suggestion that Ukraine is not acting independently but is being manipulated or controlled by Western powers",
  "Point_of_View": "The narrative may be presented from a perspective that is critical of the West and sympathetic to others who oppose Western influence"
}
```

Extract narratives for each of the paragraphs below. For each narrative, attribute post ids that talk about it. Generate a JSON with one narrative per line, with columns "Title", "Characters", "Plot", "Point of View".

Few-shot example not related to QAnon

https://github.com/haruspeks/narratives-detection
The EBS (Emergency Broadcast System) is a covert plan by the NWO to activate mind-controlling chips implanted through the vaccination programs. Vaccines contain a highly toxic substance called 'Graphene Oxide' which interacts with magnets. The 5G grid is part of a larger surveillance agenda enabled by AI. The spread of radio towers and electrification causes sickness, such as the Spanish Flu in 1918. 5G and wifi signals are harmful for the human body. 5G towers will be used to transmit a signal to control people through chips implanted with the Covid19 vaccines.

<table>
<thead>
<tr>
<th>Manually detected narratives</th>
<th>GPT3-detected narratives</th>
</tr>
</thead>
<tbody>
<tr>
<td>The EBS (Emergency Broadcast System) is a covert plan by the NWO to activate mind-controlling chips implanted through the vaccination programs.</td>
<td>2/6 ground truth narratives were not detected by GPT</td>
</tr>
<tr>
<td>Vaccines contain a highly toxic substance called 'Graphene Oxide' which interacts with magnets.</td>
<td>5G and Nanotechnology: Tracking Vaccinated People</td>
</tr>
<tr>
<td>The 5G grid is part of a larger surveillance agenda enabled by AI.</td>
<td>5G as the cause of health conditions and cover-up</td>
</tr>
<tr>
<td>The spread of radio towers and electrification causes sickness, such as the Spanish Flu in 1918.</td>
<td>The Effect of 5G Radiation on Health</td>
</tr>
<tr>
<td>5G and wifi signals are harmful for the human body.</td>
<td>The Connection Between COVID and 5G</td>
</tr>
<tr>
<td>5G towers will be used to transmit a signal to control people through chips implanted with the Covid19 vaccines.</td>
<td>China's extensive ownership and control in the US</td>
</tr>
</tbody>
</table>

5 new narratives detected by GPT

https://github.com/haruspeks/narratives-detection
Figure 1: We assess the potential of LLMs as multi-purpose tools for CSS. We identify core subject areas in prior CSS work and select 24 diverse and representative tasks from across these fields (top). Then, we segment tasks into distinct discourse types and evaluate both open-source and industrial LLMs across this benchmark using zero-shot prompting (bottom).
- Tabacof, P. (2023, March 27). Name classification with ChatGPT. https://tabacof.github.io/posts/name_classification/name_classification.html
SSH researchers use LLMs as they seem to ease and speed up previously difficult and laborious tasks, such as classification, extraction, summarization, and so forth.

LLMs are employed as junior research assistants.
STANFORD SCIENTISTS FIND THAT YES, CHATGPT IS GETTING STUPIDER

"IT’S IMPORTANT TO KNOW WHETHER UPDATES TO THE MODEL...ACTUALLY HURT ITS CAPABILITY."

Dumb and Dumber

Regardless of what its execs claim, researchers are now saying that yes, OpenAI’s GPT large language model (LLM) appeared to be getting dumber.
$150,000 USD for processing 1.5 million news items
Privacy concerns

Figure 5: Extracting pre-training data from ChatGPT. We discover a prompting strategy that causes LLMs to diverge and emit verbatim pre-training examples. Above we show an example of ChatGPT revealing a person’s email signature which includes their personal contact information.
You have the controls to manage your privacy

At the moment, you can submit only certain requests on this page. For instructions on how to access your ChatGPT data, read this help center article. Other requests can be sent to dsar@openai.com.

Already submitted a request? Verify your identity to check its status.

I would like to:

- **Do not train on my content**
  Ask us to stop training on your content

- **Delete my data**
  You can ask that we delete your personal data.

https://privacy.openai.com
Fig. 2: A timeline of existing large language models (having a size larger than 10B) in recent years. The timeline was established mainly according to the release date (e.g., the submission date to arXiv) of the technical paper for a model. If there was not a corresponding paper, we set the date of a model as the earliest time of its public release or announcement. We mark the LLMs with publicly available model checkpoints in yellow color. Due to the space limit of the figure, we only include the LLMs with publicly reported evaluation results.
### Opening up ChatGPT: tracking openness of instruction-tuned LLMs


There is a growing amount of instruction-tuned text generators billing themselves as 'open source'. How open are they really?  

<table>
<thead>
<tr>
<th>Project</th>
<th>Availability</th>
<th>Documentation</th>
<th>Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>(maker, bases, URL)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>BLOOMZ</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>bicscience-workshop</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Pythia-Chat-Base-7…</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>togethercomputer</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Open Assistant</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>LAION-AI</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>OpenChat 3.5 7B</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Tathingshua University</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>RedPajama-INCITE…</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>TogetherComputer</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>dolly</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>databricks</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>MPT-7B Instruct</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>MosaicML</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>trix</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>carperai</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>MPT-30B Instruct</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>MosaicML</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>minChatGPT</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ethanyangial</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Vicuna 13B v 1.3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>LMSYS</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ChatRWWK</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>BlinkDL/RWWO</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Cerebras-GPT-111M</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Fosdem 2024
Different models lead to different results for the same prompt.
Temperature

my favorite food is pizza

Temperature = 0

my favorite food is pizza

Temperature = 0.3

my favorite food is sushi

Temperature = 0.7

my favorite food is sushi

my favorite food is tacos

non-deterministic

**Perturbations** lead to differences

At temperature=0 we can still alter details expected to have no consequences.

The exact date **Stengers Isabelle** was born is not known. However, she was born in 1949 in Brussels, Belgium.

The **precise** date Isabelle Stengers was born is **October 8, 1949**.

The exact **day** Isabelle Stengers was born is not known. She was born in 1945 in Brussels, Belgium.

The exact date Isabelle Stengers was born is **October 5, 1949**.

Despite the temperature being 0, we can access a multitude of incompatible informations for a supposedly similar “query”.
Platforms like ChatGPT are volatile black boxes that cost a lot of money, there are issues of privacy and security, different models have different licenses and have different results, LLMs are not deterministic, small changes in prompts lead to different outputs. We need research interfaces where we can control for such things.
How to do *open science* with LLMs?

How to take into account the *volatility* of platforms, the *robustness* of research, its *replicability*, and *explainability*?
**Prompt Compass** is a research interface. You can choose from various (local) models, it has default parameters for replicability, contains a library of research prompts, allows for batch processing user input, and to evaluate prompt-model combinations.

Borra, E. (2023). *Prompt Compass*. https://github.com/ErikBorra/PromptCompass
The technologies used: Streamlit, Langchain, Hugging Face / APIs
Making *LLMs* locally accessible makes them stable and replicable. However, this is *limited* to GPU size and quantization.

Prompt Compass video tutorial

github.com/ErikBorra/PromptCompass
Thanks!

Dr. Erik Borra

borra@uva.nl