OSET INSTITUTE
STATEMENT

Considerations for Responsible Use of Generative Artificial Intelligence Technology in Support of US Elections Administration

July 7, 2023 Palo Alto, CA. The implications of artificial/augmented intelligence in elections, particularly for the 2024 U.S. national election cycle and elections in the EU, India, and Mexico portend serious consequences from chaotic disruption to the stability of democracy. A simple search of this topic is already yielding considerable results.[1] Accordingly, today the OSET Institute sets forth five principles for responsible research and development of generative AI in support of US elections administration.

Rapid advances in artificial intelligence (AI) technology pose new threats to election security, not limited to creation and amplification of disinformation and misinformation. The pace of development of new AI applications means there will be near term threats that we cannot yet anticipate for the 2024 election. At the same time, there is scope for natural-language AI techniques to create services that election officials can use to greatly amplify their very limited capabilities for voter communication, with dual-use not only for voter benefit, but also for combating disinformation.

The benefits and challenges in such use are similar to opportunities in many other settings for a limited, "domain specific" natural-language interactive system focused narrowly on one topic area. However, the risks and consequences are particularly high in the context of elections. There is a growing set of examples of naive, simplistic uses of general-purpose large-language model (LLM) technology for "advice chatbots" that provide incorrect information (*technically called "hallucinations"*) ranging from fabricated, to counterfactual, to pernicious and harmful. Such naive application of dual-use voter information services could be profoundly harmful to trust in election administration and election results, with significant downside risks of hallucinations that provide inaccurate information that impedes voter participation, fuels distrust, or even validates current misinformation or amplifies new disinformation.

Given these risks and stakes, it is *essential* to begin investigating responsible use of generative AI for narrow domain-specific applications, both generally and specifically. Generally, there are requirements for accurate domain-specific usage that are common across domains, not limited to elections. Specifically, election-related usage has particularly stringent requirements including, but not limited to, identification of disinformation and misinformation specific to elections.

Proposed Guiding Principles

There are five principles, if not mandates, that the OSET Institute is calling for in proceeding with any research, let alone development of AI applications in elections administration:

1. **No Use of General LLM Technology**
    - Use of generally available LLM technology for natural language interaction (NLI) should be strictly avoided; the downside risks mentioned above are unacceptable in elections administration.

[1] See: https://www.google.com/search?q=%22artificial+intelligence%22+elections

2. Use of Domain Specific Techniques

- As with many domain-specific applications, construction of NLI systems must be based on a narrow "foundation model" (e.g., *Titan*, *or the smallest size of LLaMA*[2]), with training on **authoritative** domain-specific information.

- Fortunately, for elections, authoritative training data is straightforwardly available for human collection and curation: state election law; state and local election offices' procedural and training materials and web content; official government documents and web content identifying prior election mis/disinformation. Training data must also include content that comprised election mis/disinformation, tagged as falsehoods.

- Also specific to elections, training data of state-specific origin must be tagged to note origin in that specific state, so that a trained NLI tool can take into account its user's origin, and provide information that is specific to the state.

3. Extensive and Bleeding-edge Use of Human Guided Training Techniques and Toolsets

- Human-guided interactive training will need to be conducted by people with elections expertise (*rather than by those with general AI training experience*) with the ability to grade AI output, rank lists of output candidates, re-write output to provide example canonical output, all based expertise in election administration and election dis/misinformation.

- Trainers should be assisted by the use of multiple different LLMs that can serve as "critics" of the primary LLM's output, make suggestions for trainers, and for defining "guardrails" for primary LLM operation.

- As a consequence of these requirements, an effective and repeatable training process will require support by training tools and environments, which are currently only in nascent form.

- Election information requires a high degree of user trust, therefore, NLI output needs to include selected references of source material — a form of "AI evidence" that is also a practice in nascent form; and since trainers must review evidence as well, this is also a relevant part of requirements for trainer toolset support.

- "Personality" tuning and guardrails will also be required, and the results assessed. An election NLI should *not* give the appearance of impersonating a human, being creative or engaging, but rather should be a source for concise, narrow, and authoritative responses only to relevant user input.

[2] LLaMA (**L**arge **LA**nguage **M**odel **A**nalysis) is a collection of state-of-the-art foundation language models ranging from 7B to 65B parameters. LLaMA is an auto-regressive language model, built on the transformer architecture (see: https://en.wikipedia.org/wiki/Transformer_(machine_learning_model)). Like other prominent language models, LLaMA functions by taking a sequence of words as input and predicting the next word, recursively generating text. What sets LLaMA apart is its training on a publicly available wide array of text data encompassing numerous spoken languages.

4. Operational Controls and Logging for Ongoing Model Refinement

- Any operational NLI must include:
  i. Guardrails, including those based on real time operation of AI critics;
  ii. Other guardrails developed during the training process;
  iii. Detailed logging of requests, responses, and guardrail operations;
  iv. Interface for user feedback and logging of feedback; and
  v. Support for trainers to review logs and use them for continuing model refinement.

5. Closed Pilots Before Public Use

- Even with the best training processes, success is hardly guaranteed, and the results of early efforts may still yield NLIs that can be manipulated into guardrail violations or into hallucinations.

- As a result, any practical use of an elections NLI *must* include a **closed** pilot phase during a real election, where pilot participants can test the utility and accuracy of the NLI with respect to current events, while ensuring that there is no public general access to a system that is not "*ready for prime time*."

- Pilot activity must include periodic refinement work with trainers reviewing logs and feedback, and further model refinement.

This topic and its implications will achieve a fevered-pitch of media coverage, hyperbole, and wide-ranging commentary with calls for (Congressional) action. It is essential that the rise of AI be taken seriously, with intellectual honesty and void of commercial or political agenda. We have offered the foregoing as a nascent start to proceeding with judicious caution and expediency ahead of 2024.

That observed, we view the likely aspirations and intentions of those giddy about AI in elections not much different than the misguided dreams of iVoting entrepreneurs—we need responsible research, but let's slow our roll toward any notion that the next big thing in election technology innovation is plying it with AI and pushing it into market.

Whatever research the OSET Institute performs will, of course, be fully transparent and built on public technology. Regardless, we have modest expectations of what exactly can be produced in time for the 2024 election cycle, especially the final three months. Yet, stay tuned, because we are seriously assessing what can be (*responsibly*) accomplished. Here is the limit of "guestimation" on what: something non-zero.

For more information, please reach out to hello at osetinstitute dot org.

# # # # #