Penn Media, Technology and Democracy

LLM Civic Discourse: A Primer

Large Language Models (LLMs) are playing an increasingly substantial role in the information ecosystem and in civic discourse—that is, how LLMs discuss political topics, refer to politicians, and relate election information. Through chatbots, search engines, writing assistants, social media websites, and other web applications, LLMs likely regularly reach billions of people. We see this issue as a critical new frontier of political discourse, akin to social media two decades ago. There is evidence that LLMs are influential and persuasive on key political issues, and there is every reason to expect their influence on civic discourse to grow.

An Emerging Research Field

An emerging body of research demonstrates the impact of LLMs on civic discourse. On a positive note, LLMs are able to reduce false election denial beliefs through targeted conversations. Although standard models performed poorly, LLMs with access to curated high-quality context may enable automated fact-checking. There is early evidence that LLMs may deepen polarization, but have a countervailing impact by lowering hostility and toxicity.

However, LLMs also demonstrate clear biases. For instance, identity-related speech is more likely to be incorrectly suppressed than other speech. More generally, LLMs display ideology that is similar, both geographically and by language used, to their developers. While that bias might result from more subtle development choices, longitudinal monitoring has demonstrated that LLM companies are making unannounced content moderation choices on politically salient topics. Research also shows that optimizing LLMs for "truthfulness" can result in a left-leaning bias, suggesting that attempts to make "unbiased" models could come at a cost to truthfulness.

LLMs may also be aiding authoritarian information control—prompts written in languages of countries with less press freedom tend to exhibit stronger pro-regime bias (source embargoed). Research has demonstrated clear concerns about censorship, for instance, with higher refusal rates on political topics from China-originating models than from those developed outside China.

Piling onto an already challenging market for journalism, LLMs built into search engines are diverting viewers and revenue from news websites, even as they depend on these sources for current information. A poll from Pew Research found that encountering an AI summary led to a 50% reduction in clickthrough rates. There is some evidence that LLMs may draw attention to dominant international news sites, rather than smaller, independent, or local journalism. (*Note: There is more research documented in the accompanying annotated bibliography).*

That LLMs are updating and evolving poses a challenge to future research—developing more comprehensive longitudinal monitoring of LLMs today, especially when resulting in open-access datasets, is essential to answering the key research questions of tomorrow. Large-scale studies on how users interact with LLMs are also critical, in order to guide the predetermined prompts used for longitudinal monitoring efforts: automated monitoring will need to mirror real human interactions to produce meaningful results. It is also clear that the cost of large-scale LLM

studies will be problematic and at times prohibitive for researchers, warranting consideration for cross-institutional research collaborations.

The Societal Impact of LLM Civic Discourse

All told, the emerging research already demonstrates that LLMs are going to play a societally impactful role in the information ecosystem, justifying a more considered and thorough approach to monitoring and evaluating how LLMs shape civic discourse. Comprehensive and longitudinal studies of LLMs may inform market choices—which LLMs businesses choose for deploying in applications, and which LLMs consumers choose to engage with news and politics.

This research also acts as an essential transparency mechanism, encouraging disclosures of policies and moderation choices by LLM providers, and identifying undisclosed policies—notably, one prominent index shows meaningful declines in LLM transparency. The political outputs of LLMs are not accidental nor unconnected from the decisions of their developers. The LLM companies are making distinct choices based on different normative goals. For instance, Anthropic states that it is seeking "political even handedness" through its (open-source) automated evaluation framework for Claude models. Alternatively, OpenAI states that ChatGPT "shouldn't have political bias in any direction," and operationalizes this primarily through seeking "objectivity." Reporting suggests that Google's Gemini has been far more reticent in responding to prompts about political figures and elections, relative to other LLMs.

These normative goals are critical for setting the future direction of LLMs, and academic researchers and civil society organizations should play a role—already, some have argued that "neutrality" is an ill-posed and potentially harmful goal for LLMs. Model developers must also contend with the tradeoff between free expression and safety concerns such as factual accuracy and discriminatory speech. This is especially challenging to apply at scale because countries vary greatly on what constitutes permissible speech.

These normative goals are reflected through intentional design of technical interventions, before the training, during the training, and during the deployment of LLMs. This includes the selection and curation of training data; the model architecture and optimization specifications; post-training including fine-tuning and reinforcement learning through human feedback; the design of system prompts; and the technical infrastructure into which an LLM is deployed.

A range of policy interventions touch on LLM civic discourse, and it is reasonable to expect increasing attention from policymakers. An alliance of NGOs and media companies have argued to European regulators that Google's AI answers violate key provisions of the European Union's (EU) Digital Services Act. The EU AI Act also covers some LLMs, and its code of practice may enforce transparency requirements and risk-mitigation that are relevant to civic discourse, especially in preventing nefarious use of LLMs to undermine democratic processes. So far, notable state laws in California and New York on frontier AI have not touched on these impacts, primarily focused on hard security risks. Yet, this may change. Leading civil society figures have raised the scenario that future laws will seek to ban LLMs from discussing controversial political topics—such as abortion.