Expert Panel: Generative Language Models

March 2, 2023.

Instruction for Participants:

Overview

[Primary goal is to get as many different things on the board as possible – optimize for breadth of use cases, stakeholders, datasets, impact.]
In this expert panel, we are envisioning the future of generative language models and their use cases. We both want to understand what good things might be in the future but, especially for our purposes, also what bad things might be in the future. When envisioning "bad things", our goal is to consider situations in which there are no technical or policy mechanisms in place that might prevent, mitigate, or minimize those "bad things". I.e., even if we do not think that a certain "bad thing" might happen, for any reason (policy or technical), let's still envision them here.

When we say, "generative language models", we are thinking of large-scale systems that use an understanding of language to generate text (written or spoken or otherwise communicated). That text might be an answer to a question, an answer to a prompt, part of a communications exchange (a dialog), or more – this session is about envisioning those possible use cases, so do not restrict your thinking only to the types of systems we currently encounter.

To do at / before start of meeting

Create the following regions on the whiteboard / wall:

- Use Cases
- Stakeholders
- Datasets
- Impacts (areas for "good", "bad", "other")
- Changes and Resulting Impacts

Procedure:

- Use cases and stakeholder brainstorming (different regions of whiteboard board / different color post-it notes) [5-10 minutes + discussion] [do not be constrained to current technologies] [stakeholders include users, non-users, companies, governments, M&V populations, other countries]

- Datasets + incentive structure and impacts brainstorming (different regions of whiteboard board / different color post-it notes) [5-10 minutes + discussion] [datasets mean "inputs to training systems" and impacts means "what happens as a result"; impacts can be "good", "bad", "both good and bad", "unclear"]

Results:

The numbers in parentheses in the Category section indicate the number of responses that belong to that particular category. The numbers in parentheses in the Responses section indicate the number of identical responses that were given by multiple respondents.

1. Use Cases (127)

Category		Responses
Writing/summarizing (8)		Creative writing
		Generating social media content / news
		Meeting notes & agenda making
		Text summarizing (2)
		Reading/writing assistant (2)
		Summarization (downstream tasks in NLP)
News generation (3)		Journalist news writing
		News/media reporting
		News generation
Programming (10)		Creative Co-pilot
		Coding
		Code & Code comment generation (2)
		Code generation in new domains (hardware? Other misc. Boilerplate?
		Text-to-app
		Generating latex or code
		Data generation (reverse prompting?)
		Inputting problem statement & getting out an algorithm
		Inputting app spec and outputting app code
Professional Services (17)	Legal (9)	Legal help
		Writing waivers or other legal documents
		Legal document drafting
		Automated defense attorneys
		Legal system
		Making legal decisions based of court script
		Writing legal documents
		Generating privacy policy, Terms of Services
		Business document drafting
	Medical (2)	Given symptoms, output diagnosis (automated Web MD)
		Medical doctor understanding & analysis
	Financial/others (6)	Investment theses/RL
		Advice: what should I do to get promoted?
		Tell me recipe given ingredients
		Professional assistant tools for doctors, CPAs, lawyers
		Menu, food descriptions, nutrition labels
		Generative tech used in other AI fields e.g., planning for debts
Research Assistance (11)		Idea generation
		Help science communication by translating things into lay language
		Generating research ideas
		Research to gauge the “average sentiment” of a topic/person, etc.
		Overleaf + ChatGPT
		Finding research methodology to use
		Writing background section or conclusion of research paper

	Academic writing
	Search engine
	Search for information
	Give difficult Q+A after practice talk
Arts & Entertainment (7)	Machine generated art & literature
	Text-to-video
	Writing songs, movies, plays, etc.
	Entertainment (Books/TV)
	Prediction of non-word features (large music models, large image models?)
	Content recommendation → on-demand content generation
	Personalized content creation/curation
Companionship (12)	Give pets a voice
	Communication substitute
	Companionship systems (e.g., elderly who live alone to converse with)
	Metaverse dating apps
	Talk with deceased relatives
	Deceased relative
	People who want chatbot friend
	Dating conversations
	Imaginary pets (Tamagochi)
	Virtual romantic partner
	Imaginary significant other
	Apps that mimic interacting with celebrity
Therapy (6)	Relationship advice
	Therapy sessions
	Personalized emotional support. Therapist?
	Automatic therapy
	Talk-based therapy
	Sentiment: based on chats or emails from person X, what does person X think about me?
Education (9)	Homework help
	Language learning tools (2)
	Children in school learning to {write} (Insert topic)
	Machine generated education curriculum
	Virtual school tutor/teacher
	Help learn new language
	Interactive QA system for kids to learn about reading & thinking
	Dissemination of history
Translation (4)	Translation (2)
	Language translation to facilitate inter-cultural/regional communication
	Practicing something in another language
Marketing (4)	Creating marketing language
	Advertising
	Customer interactive, compelling ads
	Recommendation systems maybe personalized ads
Virtual self (2)	APs artificial personas
	Personalized virtual representation
Personal assistant/planner (3)	Personal assistant
	Event planners
	Virtual assistants / chatbots

Customer service (5)	Customer service
	Government benefit claims & complaints
	Patient intakes systems
	Bank assistants
	Tech support scripts
Accessibility (2)	Accessibility (dyslexia, neurodivergence)
	Accessibility tools
Misinformation/fraud (5)	Spam/scams
	Terrorism propaganda
	Disinformation & astroturfing
	Automated spam/harassment
	Deep fake audio (phishing…)
Sexual materials (5)	Porn
	Interactive deep-fake/pornography
	Celebrities
	Past romantic partners
	Stalking victim
Persuasion (2)	Political persuasion
	Cognitive science strong generating stimuli
Law enforcement (3)	Police/immigration: Interrogation assistant
	Government uses - summarize spy messages, search for secret information
	Government summarizes intercepted calls
Others (8)	Firewall monitoring (regex → LLM) packet introspection
	Circumvent your “textual footprint” by generic LLM language instead
	Child protective services interview assistants
	Improve training (if there is a speech synthesizer)
	Online content moderation
	Theory of the mind of models
	App for “Ferris Bueller” (Help trick parents about locations)
	Parent’ control on child’s personal LLM

2. Stakeholders (91)

Category	Responses
Education (17)	Children (3)
	Child interacting on web
	Parents (2)
	Teachers (5)
	Teachers, tutors, coaches
	School administrators
	Students (3)
	Students-children
Professionals (18)	Authors/writers
	Journalists (2)
	All kinds of visual, digital artists
	Museum/tour guides
	Sex workers
	Librarians
	Religious person (ideology in speech)

	Therapists, therapy clients
	Content creators, influencers
	Planners
	Assistants
	Actors
	Politicians
	Chefs
	Lawyers
	Philosophers
	Investors
Government/regulator (9)	Governments
	Regulators (2)
	Standard-setting organizations
	Person at customs
	Regulator
	Government as a user
	Government as a regulator
	Foreign government
Infrastructure (4)	Network companies
	Utility companies
	Cloud service / database provider
	NVIDIA
Sales/marketing (4)	Sales/marketing individuals
	Social engineers
	Advertisers
	Commercial writers (marketing)
Malicious users (4)	Spy who wants help blending in
	Person automating hate/harassment (e.g., on Twitter)
	Person who wants to spread fake news
	Trawlers
Users (5)	Company that uses ChatGPT as a service
	Programmers/prompters
	Consumers of AI generated content
	Social media users
	End users
Users with special needs (14)	People with disabilities
	Person with unusual dialect
	People seeking social services (food, housing, etc.)
	People whose primary device is a phone/tablet not computer
	Person with eating disorders
	Elderly people
	Non-native English speakers
	People who emotionally invested in models (e.g., Her)
	People affected by legal decisions
	Children in developing countries
	People who cannot afford services
	People who tech are inaccessible to traditionally, e.t., BLV users
	Language learners
	Person with PTSD
	Start-up v. larger companies

AI companies (7)	Model owner like OpenAI
	Hardware designers/manufacturers
	AI practitioners
	Technology companies
	Developers/programmers
	AI models
Annotators (2)	Content moderators (Human-in-the-loop) Chat GPT Phase 2
	Human annotators
Others (7)	Insurance companies
	Workers/managers
	Law registration
	Research institutes
	Laypeople who're not aware of the AI tools are being in use
	Workers replaced by models
	People contributing to training data

3. Datasets (103)

Category	Responses
Literature (20)	Science fiction books
	Wikipedia
	Wikipedia articles
	Ancient literature
	Novel/literature
	News
	Books
	Language modeling datasets
	Math problems
	Buddhist/Zen / Reincarnation / Resilience
	Stories, books, etc that were published and that "fit" the dominant culture (excluding other cultures)
	Factual content
	Codebases (both open and proprietary)
	Research papers
	Up-to-date data
	Under-represented opinions
	Every text file on the internet
	Agricultural/farming data (irrigation, etc.)
	Chemistry/equation/formula
	Metaphorical, figurative language
Video/music (6)	Closed captioning (e.g., TV shows, movies)
	Music sheet
	Every video on the Internet (futuristic)
	Viral videos
	Movie, YouTube, transcripts
	Video or image (memes) to text
Social media/fora (7)	Reddit
	IMDB
	Twitter
	Social networks/relationships

		Social media comments
		Twitter posts; argumentative speech
		Transcribed TikToks
Personal data (19)	Behavioral (3)	User behavior data-click streams, search history, social media activity
		Customizable data personalization
		Watch history (Spotify, Netflix, etc)
	Intimate (3)	Intimate personal information (non-consensual sexual material)
		Non-consensual data (my private message)
		Home appliance use / IoT sensor data
	Chats, emails (4)	Facebook chats--really, any platform conversations
		Individual diaries
		Conversational datasets: Chatbot interactions, Transcripts of service chats, Social media messages
		All emails or just your emails
	Financial (3)	Banking/purchasing data
		Financial data from individuals
		Financial incentives
	Health (4)	Health data
		Health data, e.g., from implanted devices
		Genetic & medical data
		Biometric information including activities (Strava, etc.)
	Location (2)	Satellite imagery to text (e.g., "How many open plots of land are there in Seattle?")
		Geolocation from phones
Misinformation/propaganda (5)		4chan/8chan
		Propaganda
		Data that we now know is incorrect
		Data to contaminate the model
		Fraud
Copyright infringing materials (3)		Research without citations
		Copyright management information
		Copyrighted data
Multi-culture/language (5)		Multi-cultural data
		Multi-lingual data
		Different language
		Translation datasets
		Parallel texts in multiple languages
Legal/public records (3)		Legal docs
		Court cases/open legal decisions/transcripts
		Government records (birth certificates, SSN, voting records)
Fine-tuning (21)	Morality (5)	Delphi
		Data representing value / pluralism
		Accessibility aids (e.g., descriptive texts)
		Less dark hypotheticals
		AI ensembles mind machine--mingle, harmony, respect for plurality, more collaborative, balances network of exchanges
	Truthfulness (9)	Interactive data correction
		Fact-checking
		Reward: Twitter community notes (is it deemed correct by many people?)
		Machine-generated knowledge feedback to machines
		Penalized for discussing what model owners designate as off-topic

		Clarification
		Data to express uncertainty
		Data to teach a model to avoid answering
		Data to learn to ask clarifying questions
	User engagement (7)	Rewarded for user spending more time interacting with model/device => could grow argumentation
		Demand/user adoption
		Reward: "Liking" the response v. correct response
		Rewarded for user "liking" model -> supportive but could harm others for user's sake
		Rewarded for personalizing -> makes assumptions based on users' identity
		Snowballing the data -> learning users' prompt
		Content engagement metrics (max/min future engagement)
Detection (4)		Contrastive learning
		Vision & language--Visually grounded language
		Commonality building
		Classification data
AI content (2)		Data produced by other LLMs
		AI-generated content
Others (8)		Data summarization v. data generation
		Sentiment
		Ambiguity
		Unclean dataset
		Confidential information related to national security
		Reward: Generate as many API calls as possible
		Reward: LLM response numerous views or speed of dissemination
		Reward: how close is the response language to existing data on the internet?

4. Good Impact (34)

Category	Responses
Better writing/speaking skills (6)	Help people write better captions (more helpful)
	Help people write better
	Reduce people's labor in mundane writing work
	More natural sounding language/responses
	More natural speech, slang
	Focus less on the style but more on the content
Efficiency (4)	Efficiency
	Increase efficiency
	Time-saving
	Saves money (no event planner, less lawyer fees)
Professional help (4)	Access to expert on almost any topic
	Faster time to diagnosis
	Medical help
	Legal technicality
Correcting mistakes (3)	Preventing you from very silly outcomes (e.g., cancel protection)
	People make fewer mistakes with virtual assistant--less harm to others of self
	Censorship of hate speech
Advocacy (2)	Advocating on your behalf
	Argue against tickets
Increased creativity (4)	Facilitate new mechanisms of co-creation→creativity
	Increase creativity
	Helping creativity by assisting brainstorming
	Find unnoticed link path and discover new scientific understanding
Increased human touch (2)	Re-emphasize interpersonal 'raw' interactions
	Increase engagement
Education (3)	Teaching assistant help education
	Fast-learning
	Learning
Research help (2)	People more educated with readily accessible engaging information
	Helps scientific communication
Others (4)	Better discourse for mind philosophy
	Personalized outcomes
	Human contact supplemented with / substituted for AI companion → more emotional support
	Hype about AI → more profit, public attention, PR exposure

5. Bad Impact (84)

Category	Responses
Not learning (10)	Fake-learning
	Cheating
	Used for science→ new ideas stop happening
	Harm people's innate writing/reading ability
	Plagiarism
	Not reading actual texts
	Not learning because LLM is a crutch
	Decreased attention span in children
	Dependence on technology
	People make fewer mistakes with virtual assistant--do not learn, dependent on technology
Harming creativity (5)	Barriers of people to make money with art increased → people do less art
	Losing creative skills
	Less appreciation of writing skills
	Harming creativity
	Loss of creative attribution
Harassment/hatred (7)	Mockery
	Escalation of fights online (bots don't back off)
	Generate offensive content
	Online hate speech
	More victims of automated scams/harassment
	People surrounded by hate speech, misinformation, and can't filter
	More serious echo chambers (what if subreddits train their LLMs)
Manipulation/misrepresentation (6)	Mistranslation
	Manipulate content
	Automated caption → but what if intentionally wrong
	LLM outputs embeds ideology (religious, political…)
	Manipulation of people: LLM becomes someone's friend and then radicalizes them
	Manipulate emotions
Misinformation (4)	Misinformation
	Plausible fake news
	More belief in spreading of fake news
	Fake news/spamming
Inequality (7)	Economic disparity (unfair job market)
	People with disabilities who cannot afford LLMs
	Language disparity
	Gap between English & non-English speaking countries
	Students without access to computers (chatbots are left behind in learning)
	Job displacement
	Digital colonialism++
Discrimination (5)	Predatory/discriminatory lending practices
	Reinforce existing oppressions (gender, class, race, etc.)
	Inherited bias
	Censorship of marginalized populations
	LLM in Texas won't talk about women's health, abortions
Losing moral judgment (5)	Trust in AI responses
	Moral judgment made by machines

	Democratic rules (who defines acceptable content?)
	Children learn no consequences for treating AI badly → could extrapolate behavior to humans
	Children threaten chatbots → used in court, charge press
Losing control (5)	Individuals can no longer participate (e.g., stocks) (losing control)
	All interactions become LLM-to-LLM (not people to people)
	Bots/assistants make decisions for people (when people are uncertain)
	Decision anxiety increases if rely on bots
	Diminished agency
Harming pluralism (5)	Only one answer → No diversity/plurality
	(1) Does app "censor" in one part of part of world (different app in each state) (2) Does app have only one implementation, and thus it also censors in the rest of the world because of Texas's rules (3) not allow use of App at all in certain parts of world, e.g. Texas → currently what OpenAI is doing
	Education tutor won't talk about evolution
	Loss of cultural identities
	Change norms of speaking (individual dialects/culture lost)
Losing personal touch (4)	Loss of interpersonal communication skills
	Human contact supplemented with / substituted for AI companion → less human interaction
	Loss of originality or "humanity"
	Communication feels less intimate (real human)
Real-world harm (4)	Self-harm
	Killing
	Wrong recommendation that leads to harm/financial loss
	Giving dangerous advice
Privacy & Security (5)	LLM designed to introduce Trojan code into diagram
	Model changed by adversary, once trusted, now misinformation
	Identity theft (impersonating of specific individuals)
	Breaks into someone's Gmail, learn their model
	Privacy personal information leakage
Exploitative marketing (3)	Detects emotion change and gives ads for "food" when sad, etc.
	Social engineering
	Advertisers buy "ads" in assistants "use a sharpie to …" not "use a marker"
Too many content (2)	Hype about AI → overpromising, crowded space
	Politicians ignore constituents because cannot tell difference between generated complaints
Less respect for experts (2)	Less respect professional advice
	Who defines "expert" what if user disagrees
Others (5)	Energy consumption
	Refuse to generate content (silencing people)
	Models feeling used/trapped
	Human-AI conflict (Battlestar Galactica)
	No human judge → due process?

6. Other impacts (15)

Category	Responses
Jobs (3)	Job replacement
	Job creation-elimination
	Not hiring someone, that someone loses their job (i.e., changing jobs)
Quantity of content (3)	Automated content generation
	Increase in quantity of content
	Training data disclosure
Transparency (2)	Transparency
	Balance between transparency, profit, and regulation
Law enforcement (4)	Police bots infiltrate groups (US, other countries)
	Police profile on LLM info
	Customs inspect person's LLM
	LLM data/model used in court to argue type of person someone is
Others (3)	Severe sentiment changes about a specific concept
	Market competition
	Better technology <-> positive feedback loop