

Addressing Challenges in GPT Models

Ali Nourian - Negin Mohtaram

Sharif University of Technology
Department of Electrical Engineering
Prof. B. Khalaj

August 2, 2023

Table of Contents

- 1 Challenges of New Contexts
- 2 Challenges of Data Privacy
- 3 Challenges of Reliability
- 4 Reference

Challenges of New Contexts

Challenges of New Contexts

- LLMs have shown impressive capabilities in generating text based on patterns learned from vast amounts of training data.
- But, what happens when facing information that is beyond their training dataset?
- what happens when facing inaccessible private sources?

These are some challenges we are facing:

- Inventing Information
- Offering generic responses
- Prompting for clarification
- Overconfidence or misleading answers

Inventing Information

LLMs generate responses based on statistical patterns learned from training data, without truly understanding the underlying concepts.

- LLMs tends to invent information when confronted with unfamiliar topics
- they may produce plausible-sounding but inaccurate or even fictional responses in situations where they lack knowledge

Example:

Question: "What are the applications of advanced quantum physics theories?"

LLM Response: "Advanced quantum physics theories are used for time travel and teleportation."

Offering generic responses

- LLMs often default to providing generic or vague answers when they lack specific information.
- This behavior arises because they are trained on a diverse range of topics, but their responses may lack depth or specificity.

Example:

Question: "What are the symptoms of a rare medical condition called 'Xyzitosis'?"

LLM Response: "You should consult a medical professional for proper diagnosis and treatment."

Prompting for clarification

- LLMs may encounter ambiguous queries or unfamiliar terms, leading them to request clarification or additional details.
- While this can help in understanding the context better, it also indicates their limitations in comprehending nuanced information.

Example:

Question: "What does the acronym 'AI' stand for?"

LLM Response: "I'm sorry, could you please specify whether you are referring to 'Artificial Intelligence' or 'Amnesty International'?"

Overconfidence or misleading answers

- LLMs can sometimes generate responses that sound confident or accurate but are, in fact, incorrect or misleading.
- Users may perceive these responses as reliable due to the model's fluency and coherence, leading to potential misinformation.

Example:

Question: "What was the cause of the Great Fire of London in 1666?"

LLM Response: "The Great Fire of London was caused by a dragon that accidentally breathed fire on the city."

Solutions for handling the Challenges

To address these challenges and enhance the reliability of LLMs, we can employ various strategies and approaches. There are some potential solutions:

- Fact-Checking and Verification
- Contextual Clarification
- Ongoing Research Efforts like Few-shot learning
- Improved prompt engineering
- Open-domain dialogue systems

Fact-Checking and Verification

- Fact-checking and verification play a crucial role in ensuring the accuracy of information provided by LLMs.
- Users should cross-reference the responses with trusted references or consult domain experts to validate the information.
- By promoting fact-checking and critical evaluation, we can reduce the potential spread of misinformation.

Contextual Clarification

- Contextual clarification is key to improving the accuracy of LLM responses.
- By providing additional context or specifying the scope of the query, users can help LLMs generate more precise and relevant answers.
- This can be done by offering specific details, narrowing down the topic, or including relevant background information to guide the LLM's response.

Ongoing Research Efforts like Few-shot learning

The research community is actively working on enhancing LLMs' capabilities to handle new contexts and inaccessible information. Ongoing research efforts include:

- Few-shot learning: Techniques that enable LLMs to adapt to new contexts with minimal training data, allowing them to generate more accurate responses in previously unseen situations.
- Improved prompt engineering: Developing strategies to design effective prompts or queries that elicit more precise and specific responses from LLMs, guiding their generation process towards more relevant answers.
- Open-domain dialogue systems: Exploring dialogue systems that can access external knowledge sources or leverage reinforcement learning to learn from user feedback, enabling LLMs to incorporate real-time information and provide up-to-date responses.

Challenges of Data Privacy

Data Privacy Risks in LLMs

- Data collection and potential inclusion of sensitive information
- User interaction and data logging
- Data retention and associated risks
- Data security concerns
- Inference attacks and privacy breaches

Data collection and potential inclusion of sensitive information


- Sources of Training Data
- Anonymization and Privacy Preservation
- Inadvertent Inclusion of Sensitive Information
- Potential Consequences of including sensitive information



Figure: Data Collection

User interaction and data logging

- Data Logging in LLMs
- Data Retention
- Risks of Prolonged Data Retention



	Current LSN	Operation	Context	Transaction ID	SPID	Begin Time	Transaction Name
14485	000001e9 000003d0 0001	LOP_BEGIN_XACT	LCK_NULL	0000 0006f44	1154	2016/08/10 13:50:11.287	GhostCleanup Task
14486	000001e9 000003d0 0002	LOP_COMMIT_XACT	LCK_NULL	0000 0006f44	NULL	NULL	NULL
14487	000001e9 000003d8 0001	LOP_MODIFY_ROW	LCK_BOOT_PAGE	0000 00000000	NULL	NULL	NULL
14488	000001e9 000003d0 0001	LOP_BEGIN_XACT	LCK_NULL	0000 0006f45	26	2016/08/10 13:50:31.287	GhostCleanup Task
14489	000001e9 000003d0 0002	LOP_COMMIT_XACT	LCK_NULL	0000 0006f45	NULL	NULL	NULL
14490	000001e9 000003d8 0001	LOP_MODIFY_ROW	LCK_BOOT_PAGE	0000 00000000	NULL	NULL	NULL
14491	000001e9 000003d0 0001	LOP_BEGIN_XACT	LCK_NULL	0000 0006f46	26	2016/08/10 13:50:51.353	GhostCleanup Task
14492	000001e9 000003d0 0002	LOP_COMMIT_XACT	LCK_NULL	0000 0006f46	NULL	NULL	NULL
14493	000001e9 000003d8 0001	LOP_MODIFY_ROW	LCK_BOOT_PAGE	0000 00000000	NULL	NULL	NULL
14494	000001e9 00000400 0001	LOP_BEGIN_XACT	LCK_NULL	0000 0006f47	1154	2016/08/10 13:51:11.343	GhostCleanup Task
14495	000001e9 00000400 0002	LOP_COMMIT_XACT	LCK_NULL	0000 0006f47	NULL	NULL	NULL
14496	000001e9 00000408 0001	LOP_MODIFY_ROW	LCK_BOOT_PAGE	0000 00000000	NULL	NULL	NULL
14497	000001e9 00000410 0001	LOP_BEGIN_XACT	LCK_NULL	0000 0006f48	450	2016/08/10 13:51:31.337	GhostCleanup Task
14498	000001e9 00000410 0002	LOP_COMMIT_XACT	LCK_NULL	0000 0006f48	NULL	NULL	NULL
14499	000001e9 00000418 0001	LOP_MODIFY_ROW	LCK_BOOT_PAGE	0000 00000000	NULL	NULL	NULL
14500	000001e9 00000420 0001	LOP_BEGIN_XACT	LCK_NULL	0000 0006f49	26	2016/08/10 13:51:51.337	GhostCleanup Task
14501	000001e9 00000420 0002	LOP_COMMIT_XACT	LCK_NULL	0000 0006f49	NULL	NULL	NULL
14502	000001e9 00000428 0001	LOP_MODIFY_ROW	LCK_BOOT_PAGE	0000 00000000	NULL	NULL	NULL
14503	000001e9 00000430 0001	LOP_BEGIN_XACT	LCK_NULL	0000 0006f4a	22	2016/08/10 13:52:11.347	GhostCleanup Task
14504	000001e9 00000430 0002	LOP_COMMIT_XACT	LCK_NULL	0000 0006f4a	NULL	NULL	NULL
14505	000001e9 00000438 0001	LOP_MODIFY_ROW	LCK_BOOT_PAGE	0000 00000000	NULL	NULL	NULL

Figure: Data Logging

Data security concerns

- Unauthorized Access
- Data Breaches
- Encryption
- Storage & Data Transfer Security



Figure: Data Security

Privacy Enhancement Techniques in LLMs

- Anonymization and data minimization
- Differential privacy
- Federated learning
- Secure multi-party computation (MPC)
- Transparency and user control
- Ethical guidelines and regulatory frameworks

Privacy Enhancement Techniques in LLMs

Anonymization and Data Minimization

- Anonymization:
 - Remove or obfuscate personally identifiable information (PII)
 - Protect user privacy by preventing identification of individuals
 - Enables data analysis and personalized services without compromising privacy
- Data Minimization:
 - Collect and store only necessary data for LLM functioning
 - Minimize collection and retention of personal data
 - Reduces risk of privacy breaches and unauthorized access

Privacy Enhancement Techniques in LLMs

Benefits of Anonymization and Data Minimization

- Anonymization:
 - Protects user privacy by preventing identification
 - Allows LLMs to derive insights and provide personalized services
 - Mitigates risks of unauthorized access and data breaches
- Data Minimization:
 - Reduces amount of personal data collected and stored
 - Limits potential harm in case of data breaches
 - Aligns with privacy principles and regulatory requirements

Privacy Enhancement Techniques in LLMs

Differential Privacy

- Provides a rigorous mathematical framework for privacy protection
- Focuses on preserving privacy while allowing useful data analysis
- Adds noise or perturbation to query results to protect individual privacy

Privacy Enhancement Techniques in LLMs

Benefits of Differential Privacy

- Protects individual privacy by limiting the disclosure of sensitive information
- Allows for accurate data analysis while preserving privacy guarantees
- Provides a quantifiable measure of privacy protection through privacy budget

Privacy Enhancement Techniques in LLMs

Federated Learning

- Enables training of machine learning models on decentralized data
- Data remains on user devices, reducing the need for data centralization
- Aggregates model updates instead of raw data, preserving privacy
- Allows for personalized services without compromising user privacy

Privacy Enhancement Techniques in LLMs

Federated Learning in LLMs

Benefits of Federated Learning:

- Privacy Preservation:
 - User data remains on their devices, reducing privacy risks
 - Minimizes the exposure of sensitive information to third parties
- Improved Data Security:
 - Reduces the risk of data breaches and unauthorized access
 - Protects user data by keeping it decentralized and under user control
- Enhanced Personalization:
 - Allows for personalized services based on user-specific models
 - Preserves user preferences and individual characteristics

Privacy Enhancement Techniques in LLMs

Secure Multi-Party Computation (MPC)

- Allows multiple parties to jointly compute a function while keeping their inputs private
- Enables collaboration without revealing sensitive data to other parties
- Protects privacy by ensuring that no party learns more than necessary

Privacy Enhancement Techniques in LLMs

Secure Multi-Party Computation (MPC) in LLMs

Benefits of Secure Multi-Party Computation (MPC):

- Privacy-Preserving Collaboration:
 - Enables collaboration on data analysis without sharing raw data
 - Protects sensitive information by keeping it private during computation
- Data Confidentiality:
 - Ensures that no party can access or infer the inputs of other parties
 - Maintains confidentiality of individual data while performing joint computations
- Trust and Security:
 - Provides a secure framework for parties to compute collectively
 - Reduces the risk of data breaches and unauthorized access to sensitive information

Privacy Enhancement Techniques in LLMs

Transparency and User Control

- Emphasizes providing clear information to users about data collection and usage
- Gives users control over their personal data and how it is shared
- Enables informed decision-making and consent regarding privacy practices

Privacy Enhancement Techniques in LLMs

Transparency and User Control in LLMs

Benefits of Transparency and User Control:

- Informed Consent:
 - Allows users to make informed decisions about data sharing and usage
 - Enhances user trust by providing transparency in privacy practices
- User Empowerment:
 - Gives users control over their personal data and its sharing preferences
 - Enables users to customize their privacy settings according to their preferences
- Accountability and Compliance:
 - Demonstrates compliance with privacy regulations and ethical guidelines
 - Facilitates transparency in data handling practices for regulatory purposes

Privacy Enhancement Techniques in LLMs

Ethical Guidelines and Regulatory Frameworks

- Provide guidelines and rules for responsible data handling and privacy protection
- Ensure compliance with legal requirements and ethical standards
- Promote transparency, fairness, and accountability in LLMs

Privacy Enhancement Techniques in LLMs

Ethical Guidelines and Regulatory Frameworks in LLMs

Benefits of Ethical Guidelines and Regulatory Frameworks:

- User Trust and Confidence:
 - Establishes trust by demonstrating commitment to privacy protection
 - Assures users that their data will be handled responsibly and ethically
- Data Protection and Privacy:
 - Ensures compliance with legal requirements for data protection and privacy
 - Safeguards user rights and prevents unauthorized use or disclosure of personal data
- Accountability and Responsibility:
 - Holds LLM providers accountable for their data handling practices
 - Encourages responsible data governance and ethical decision-making

Challenges of Reliability

Training Approach

- ChatGPT generates responses based on learned patterns from training data.
- Pre-training: Model learns from a large dataset containing parts of the Internet.
- Fine-tuning: Narrower dataset with human reviewers following OpenAI's guidelines.

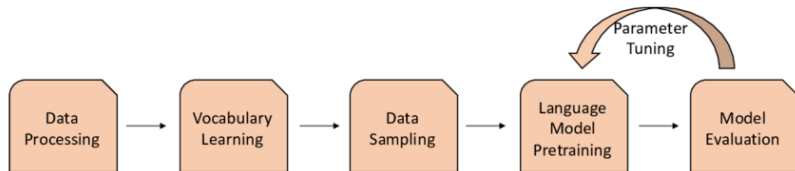


Figure: The pipeline of pre-training language models

Training Approach

Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



This data is used to fine-tune GPT-3.5 with supervised learning.



Step 2

Collect comparison data and train a reward model.

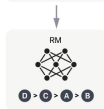
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.



The PPO model is initialized from the supervised policy.



The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



Figure: GPT-3.5 model workflow

Strategies for Increasing Reliability in ChatGPT

① Improving the Fine-Tuning Process:

- Refining guidelines for human reviewers during fine-tuning.
- Clearer instructions and additional context to reduce biases.
- Enhancing accuracy of model responses.

② Reducing Biases:

- Minimizing both obvious and subtle biases in ChatGPT's responses.
- Providing clearer instructions to reviewers regarding bias-related pitfalls.
- Valuing user feedback to identify and address biases.

③ User Customization:

- Developing an upgrade to allow easy customization of ChatGPT's behavior.
- Users can define their own values and preferences within specified bounds.
- Aligning the system with individual needs and enhancing usefulness.

Improving the Fine-Tuning Process

① Enhanced Guidelines:

- Continuous refinement of guidelines provided to human reviewers during fine-tuning.
- Clearer and more detailed instructions to reduce potential biases.
- Improving the accuracy and reliability of ChatGPT's responses.

② Iterative Feedback Loop:

- Ongoing relationship with reviewers for continuous improvement.
- Regular meetings, clarifications, and addressing questions.
- Better understanding of challenges and shared understanding of desired behavior.

③ Addressing Bias-Related Pitfalls:

- Minimizing both obvious and subtle biases in ChatGPT's responses.
- Research and development to provide clearer instructions to reviewers.
- Awareness of potential bias-related pitfalls and challenges.

④ User Feedback and External Input:

- Valuing user feedback and external input in identifying biases and issues.
- Encouraging users to report problematic outputs and provide insights.
- Refining the fine-tuning process based on feedback to improve reliability.

Reducing Bias in ChatGPT's Responses

1 Clearer Instructions to Reviewers:

- Providing explicit guidance to human reviewers about potential bias-related challenges.
- Better understanding and navigation of biases during the fine-tuning process.
- Reducing biases present in ChatGPT's responses.

2 Bias Identification and Mitigation:

- Actively researching and developing techniques to identify and mitigate biases.
- Analyzing model outputs and monitoring for potential biases.
- Iteratively improving the fine-tuning process to reduce bias-related issues.

3 External Input and Audits:

- Recognizing the importance of external input in identifying biases.
- Exploring partnerships with external organizations for third-party audits.
- Gaining diverse perspectives and ensuring a fair and reliable system.

4 User Customization and Values Alignment:

- Developing an upgrade to allow user customization of ChatGPT's behavior.
- Empowering users to define their own values and preferences within specified bounds.
- Reducing potential biases by aligning the system with individual needs.

User Customization for Personalized Experience

① User-Defined Values and Preferences:

- Upcoming upgrade enables users to define their own values and preferences.
- Shaping ChatGPT's behavior according to individual needs.
- Providing more control and flexibility to align the system with user requirements.

② Defined Bounds and Limitations:

- Certain bounds and limitations in place to ensure responsible usage.
- Preventing malicious uses or extreme customization that may lead to harmful outputs.
- Promoting a safe and reliable user experience.

③ Balancing Customization and General Usefulness:

- Striking a balance between user customization and maintaining general usefulness.
- Empowering users while ensuring ChatGPT provides helpful and accurate responses.
- Catering to diverse needs while maintaining system reliability.

Reference

References

- ❶ "Privacy in Location-Based Services: Research Challenges and Opportunities" by Yan Huang, et al.
- ❷ "Privacy in Location-Based Services: A Survey" by Xiaokui Xiao, et al.
- ❸ "The New Rules of Data Privacy" - Harvard Business Review
- ❹ "A Survey on In-context Learning"
- ❺ Large Language Models (LLMs): Challenges, Predictions, Tutorial
- ❻ "ChatGPT and large language models: what's the risk?" National Cyber Security Centre.
- ❼ "Addressing Security and Privacy Risks in Large Language Models" - OpenAI
- ❽ "Privacy in Location-Based Services: An Overview" - Claudio Bettini, et al.

Thanks! Any Questions?