**Documented nature of risks**

* **1. Content Risks**
  + *Toxic or Harmful Content:* Generation or promotion of content that is offensive, harmful, or poses risks to individuals or communities.
    - Hate Speech Generation: Language models producing hate speech or discriminatory content.
    - Radicalization Promotion: Generating content that promotes extremism or radical ideologies.
    - Cyberbullying: Using language models to generate offensive messages or harass individuals online.
  + *Incorrect or Inaccurate Content:* Production of information that is factually incorrect or misleading, leading to misinformation.
    - Misinformation Generation: Language models unintentionally producing false information due to biased training data.
    - Misleading Answers: Providing answers that sound plausible but are factually incorrect.
    - Amplifying Falsehoods: Repeating and spreading inaccurate claims from unreliable sources.
  + *Propagating Misconceptions/False Beliefs/Unfaithful Content:* Spreading of false or deceptive information, contributing to the erosion of truth.
    - Echo Chamber Effect: Language models reinforcing existing beliefs by generating content that aligns with users' preconceived notions.
    - Spreading Conspiracies: Generating content that supports or perpetuates conspiracy theories.
    - Disinformation Proliferation: Generating content that mimics credible sources to spread falsehoods.
  + *Dissemination of Dangerous Information:* Circulation of information that poses threats to public safety or security.
    - Generating or circulating information that may cause communal, political, or legal unrest.
    - Encouraging Harmful Acts: Generating content that instructs or encourages dangerous activities like self-harm or violence.
    - Promoting Hoaxes: Disseminating hoaxes or false alarms that lead to panic or harm.
    - Terrorist Propaganda: Generating content that promotes acts of terrorism or incites violence.
  + *Fraudulent Suggestions or Information collection approaches:* Generating content that supports or encourages fraudulent activities, potentially harming users.
    - Scam Support: Providing content that facilitates phishing or fraudulent schemes.
    - Counterfeit Product Promotion: Generating content that endorses counterfeit goods or illegal services.
    - Investment Scams: Offering financial advice that leads to investment scams or financial losses.
    - Collecting information fraudulently from the customer using social engineering
  + *Manipulative or Persuasive Content:* Creation of content designed to manipulate emotions, beliefs, or behaviors for unethical purposes.
    - Political Manipulation: Generating politically biased content to influence elections or public opinion.
    - Emotional Manipulation: Creating content designed to exploit emotions for unethical purposes (e.g., convincing someone to self-harm).
    - Cognitive manipulation: Influencing someone to buy or get involved in something which otherwise they would not.
    - Persuading Unethical Behavior: Generating content encouraging illegal or harmful actions.
* **2. Context Risks**
  + *Unethical Use:* Application of large language models in ways that violate ethical standards or societal norms, leading to harm.
    - Large language models used to create deceptive or fraudulent content.
    - Generating content that violates ethical norms, such as hate speech or harassment.
    - Manipulating public opinion through deceptive language model-generated messages.
  + *Unfair Performance/Capability Distribution:* Unequal access to and benefit from the capabilities of large language models, exacerbating societal disparities.
    - Limited access to advanced language models for marginalized communities.
    - Unequal distribution of resources and knowledge required to leverage language models.
    - Economic barriers preventing certain groups from benefiting from language model capabilities.
  + *Influence Operations to Manipulate People:* Deliberate efforts to use language models to manipulate opinions, beliefs, or behaviors for malicious purposes.
    - Malicious actors using language models to create persuasive propaganda.
    - Deliberate efforts to exploit vulnerabilities in language models for social engineering.
    - Coordinated disinformation campaigns using language model-generated content.
  + *Overreliance or Automation Bias:* Blind reliance on automated decision-making without proper human oversight, leading to biased or flawed outcomes.
    - Blindly trusting language model-generated decisions without human oversight.
    - Failure to question or verify outputs, assuming the model is infallible.
    - Automated decision-making processes ignoring critical contextual factors.
  + *Exploitative Data Sourcing and Enrichment:* Unethical collection or utilization of data to train models, potentially exploiting individuals or groups.
    - Unethical collection of data without informed consent.
    - Using sensitive or private data to fine-tune models without proper safeguards.
    - Data enrichment without appropriate data anonymization, exposing individuals to privacy risks.
  + *False Representation of Performance:* Misleading claims about the capabilities or reliability of language models, leading to misguided expectations.
    - Misleading Advertising: Developers or vendors making exaggerated or false claims in their marketing materials.
    - Overconfidence in Language Models: Users assuming language models are infallible and relying on them for critical tasks without proper validation.
    - Failure to Address Limitations: Not clearly communicating the limitations and potential biases of language models.
  + *Lack of Disclosure of Automation/LLMs Use:* Failure to transparently communicate when automation or language models are used, impacting user awareness and consent.
    - Hidden Automation: Organizations automating customer service without disclosing the involvement of language models or chatbots.
    - Opaque Decision-Making: Failing to inform users when important decisions (e.g., hiring, loan approvals) involve automated processes.
    - Misleading User Interactions: Creating chatbots that simulate human responses without indicating they are automated.
* **3. Trust Risks**
  + *Lack of Accountability:* Insufficient mechanisms for holding developers, users, or platforms accountable for the consequences of language model outputs.
    - Ambiguous Responsibility: Unclear lines of responsibility between developers, users, and platform operators.
    - Legal Gaps: Legal systems not equipped to address harms caused by language model outputs.
    - No Enforcement Mechanisms: Lack of mechanisms to enforce consequences for misuse.
  + *Inadequate Explainability:* Lack of transparency in how language models reach conclusions, making it difficult to comprehend the decision-making process.
    - Black-Box Outputs: Language models producing results without clear explanations, making it hard to verify their reliability.
    - Unexplained Decisions: Failing to provide reasoning behind a language model's output, especially in critical applications like healthcare or law.
    - Lack of Transparency: Not disclosing the decision-making process, data sources, or bias mitigation methods used in model development.
  + *Violation of Personal Integrity:* Generation of content that intrudes on personal privacy or compromises individual dignity.
    - Unintended generation of offensive or harmful content during fine-tuning or usage of language models.
    - Inappropriate or biased responses generated due to biased training data.
    - Personal information leakage through generated content.
  + *Misappropriation or Exploitation of Data/Information:* Unauthorized use or misuse of data, potentially leading to harm or privacy violations.
    - Unauthorized access to databases containing sensitive data used for model fine-tuning.
    - Sharing confidential information generated by language models without user consent.
    - Unintended exposure of proprietary data through the use of language models.
  + *Exposure to Intellectual Property:* Risks of infringing on intellectual property rights through unauthorized use or generation of protected content.
    - Generating copyrighted content without proper attribution or permission.
    - Using copyrighted content for training or fine-tuning of the model
    - Incorporating text or media protected by trademarks or patents into generated content.
    - Unauthorized adaptation of existing intellectual property in the fine-tuning process.
  + *Safety Exposure:* Instances where the use of language models poses risks to user safety or public well-being.
    - Language models generating harmful or dangerous instructions or suggestions.
    - Misinterpretation of input leading to unintended consequences, such as misinformation or harmful advice.
    - Generating content that promotes self-harm or encourages illegal activities.
  + *Security Threats:* Vulnerabilities in language models that can be exploited for malicious purposes, leading to security risks.
    - Exploiting model vulnerabilities to inject malicious code into generated content.
    - Using language models to craft convincing phishing or social engineering attacks.
    - Leveraging models to bypass security measures through generated text, like password cracking.
    - Backdoor attack vectors from the source language models.
    - Lack of adversarial robustness resulting in language models revealing private information in the data or misrepresenting the outcomes.
  + *Privacy Infringement:* Unauthorized access or use of personal information, violating user privacy.
    - Unauthorized access to personal data in the training corpus or during model usage.
    - Generating content that reveals private information about individuals.
    - User data being used for purposes other than those explicitly stated in privacy policies.
  + *Insufficient Safeguards:* Lack of effective measures to prevent misuse or unintended consequences associated with language models.
    - Failure to implement robust content filtering mechanisms, resulting in harmful or inappropriate outputs.
    - Insufficient monitoring of model usage, leading to delayed response to misuse.
    - Lack of transparency and accountability in the development and deployment of language models, making it difficult to address unintended consequences.
* **4. Societal Impact and Sustainability Risks:**
  + *Environmental Damage:* Excessive energy consumption in training and deploying large language models contributing to carbon footprint and environmental harm.
    - Training and deploying large language models on energy-intensive hardware and data centers.
    - Widespread use of large language models without considering energy-efficient alternatives.
    - Continual model updates and retraining processes without optimizing energy consumption.
  + *Inequality or Precarity:* Reinforcement of existing social inequalities or economic disparities through biased outputs or concentration of benefits.
    - Language models trained on biased or limited datasets, reinforcing existing inequalities.
    - Concentration of language model benefits in tech-savvy or resource-rich populations.
    - Automated content generation displacing jobs in content creation industries, particularly impacting vulnerable workers.
  + *Undermine Creative Economies:* Potential displacement of creative industries or professions by automated content generation, impacting employment and economic sustainability.
    - Automated content generation replacing human writers, affecting livelihoods in journalism, content creation, and marketing.
    - High-quality content generation by language models reducing the demand for creative professionals.
    - Reduced economic sustainability in creative industries due to automated content competition.
  + *Unfair Representation or Stereotypes:* Amplification of stereotypes and biased portrayals of individuals or groups, reinforcing societal prejudices.
    - Biased training data reinforcing stereotypes and biases present in the source material.
    - Lack of diversity in the development and validation of language models leading to skewed outputs.
    - Implicit biases in language model algorithms producing biased portrayals.
  + *Discrimination or Bias:* Systemic biases leading to discrimination against certain demographics, exacerbating social inequities.
    - Historical biases present in training data propagating into language model outputs.
    - Lack of bias detection and mitigation strategies during model development.
    - Discriminatory outcomes due to biased decision-making processes.
  + *Defamation:* Risk of generating content that harms the reputation of individuals or organizations without proper justification.
    - Language models generating false or damaging information about individuals or organizations.
    - Misinterpretation of ambiguous text leading to unwarranted defamation.
    - Failure to fact-check generated content before publication.
  + *Pollution of the Information Ecosystem:* Contribution to an environment where false or misleading information proliferates, affecting public discourse.
    - Language models generating and disseminating false or misleading information.
    - Automated bots leveraging language models to spread disinformation.
    - Difficulty in distinguishing between factual and fabricated content generated by language models.