05
Practice Session

Session Prompt

- **Scenario:** An international media company is looking to cut costs and increase efficiency by procuring a custom LLM solution from a vendor to streamline reporting. They want their journalists to **use the LLM for research, analyzing and summarizing information, and writing and editing drafts of articles**. However, they know there may be some human rights risks, so they have hired you to conduct a human rights assessment.
- **Use Case: A journalist is using the LLM for the first time as part of reporting on Covid-19**
- **Prompt: Identify some of the human rights risks associated with developing and using an LLM for this use case, and make recommendations for how they might be mitigated**. Think about recommendations for:
  - **the engineers** developing the custom LLM
  - **the media organization** procuring it and directing employees to use it
  - **the specific journalist** using it to report on Covid-19
  - **the broader ecosystem** (e.g. on public policy or industry collaboration)

Instructions

1. **Assemble into breakout groups of ~5 people**. Identify a group scribe and someone who will present your results
2. **Open the practice session materials on Google Drive here: https://bit.ly/3AU5APQ** It contains 1) a slide deck with instructions and materials to help you, and 2) an assessment spreadsheet. Please create a new copy of the spreadsheet for your group.
3. **Step 1: Use the long list of human rights to brainstorm potential human rights risks** associated with the prompt. Write down at least 5 risks in the assessment table (spreadsheet) with the corresponding human right. Try to be specific!
4. **Step 2: Use the severity criteria to assess scope, scale, and remediability** for each of your risks. The assessment table will spit out an overall severity score that you can use to compare the risks you identified and gauge whether your assessment feels right.
5. **Step 3: Write down recommendations to address the risks** you identified in the corresponding spot on the spreadsheet.

Long List of Human Rights

- ❑ Right to equality and non-discrimination
- ❑ Right to life, liberty, and personal security
- ❑ Freedom from slavery
- ❑ Freedom from torture and degrading treatment
- ❑ Due process and fair trial rights
- ❑ Freedom from arbitrary arrest and exile
- ❑ Right to privacy
- ❑ Freedom of movement
- ❑ Right to asylum
- ❑ Right to a nationality and the freedom to change nationality
- ❑ Right to marriage and family
- ❑ Right to own property
- ❑ Freedom of thought
- ❑ Freedom of religion and belief
- ❑ Right to remedy
- ❑ Freedom of opinion, expression, and access to information
- ❑ Right of peaceful assembly and association
- ❑ Right to political participation
- ❑ Right to social security
- ❑ Labor Rights (e.g. safe working conditions, adequate remuneration, right to join unions)
- ❑ Right to rest and leisure
- ❑ Right to adequate living standards
- ❑ Right to health
- ❑ Right to education
- ❑ Right to participate in the cultural life of the community
- ❑ Right to benefit from scientific advancement
- ❑ Right to internet access
- ❑ Right to a healthy environment
- ❑ Disability rights (e.g. right to accessibility)
- ❑ Child Rights

HRA Spreadsheet

Human Right	Risk	Scope	Scale	Remediability	Severity Score
e.g. Right to Privacy	e.g. The custom LLM tool could output information about sensitive sources if that was included in the data from the media organization used to fine tune it	Small / Medium / Large	Less Serious / Somewhat Serious / Very Serious	Possibly Remediable / Rarely Remediable / Not Remediable	Lowest = 3, Highest = 9
		Small	Very Serious	Rarely Remedia...	6
					0
					0
					0
					0
					0
					0
					0
					0
					0
					0
Recommendations					
e.g. The media organization should ensure no sensitive material is included in the data shared with the developer to fine tune the LLM					

Severity Assessment Criteria

Criteria	Levels		
Scope How many people are (or could be) affected by the adverse impact?	**Small** Minority range of the relevant population impacted.	**Medium** Over half of the relevant population impacted	**Large** Significant or all of the relevant population impacted.
Scale How serious are the impacts (or could they be) for affected individuals?	**Less Serious** Associated with indirect and/or minimal to moderate adverse impacts on physical, mental, civic, or material well being.	**Somewhat Serious** Associated with direct and/or serious adverse impacts on physical, mental, civic, or material well being.	**Very Serious** Associated with lasting adverse impacts on physical, mental, civic, or material well being.
Remediability Can a remedy restore affected individuals to the same or equivalent position before the adverse impact?	**Possibly Remediable** There is possible to a remedy that would return those affected to the same or equivalent position before the adverse impact occurred.	**Rarely Remediable** Remedy can rarely return those affected to the same or equivalent position before the adverse impact occurred.	**Not Remediable** Remedy will not return those affected to the same or equivalent position before the adverse impact occurred.

Human Right	Risk	Scope	Scale	Remediability	Severity Score
e.g. Right to Privacy	e.g. The custom LLM tool could output information about sensitive sources if that was included in the data from the media organization used to fine tune it	Small / Medium / Large	Less Serious / Somewhat Serious / Very Serious	Possibly Remediable / Rarely Remediable / Not Remediable	Lowest = 3, Highest = 9
		Small	Very Serious	Rarely Remediable	6
Recommendations					
e.g. The media organization should ensure no sensitive material is included in the data shared with the developer to fine tune the LLM					