ProMPT

Professional Model Prompting Tool

Using AI to Rebuild War-Torn Nations

334681, LOPES SUESDEK ROCHA, **SOFIA**
334062, MERELLI, **LEONARDO**
347839, NYKVIST, ROLF **FILIP** SEBASTIAN
328173, PARICHEHREHTEROUJENI, **MOHAMAD**

About ProMPT

System description: decision support for applying AI solutions to post-war problems raising awareness of AI applications by suggesting potential risks and potential mitigations for those risks.

Project website: https://github.com/Sofia-ME/ProMPT

Responsible AI blind spots

Unnoticed biases and potential blind spots

19%

6 actions to take

12 taken

Actions to take

● critical ● pressing ● inapplicable ● covered

Uses

Oversight

Continuously monitor metrics and utilize guardrails or rollbacks to ensure the system's output stays within a desired range (suggested).

Team

Ensure team diversity (suggested).

Train team members on ethical values and regulations (suggested).

Harms

Develop strategies to mitigate identified harms or risks for each intended use (suggested).

Data

System

Review the code for reliability (suggested).

Document the security of all system components in consultation with experts.

Actions to take

System

System security

Document the security of all system components in consultation with experts.

Think about: Periodic review by experts

Harms

Harm resolution

Develop strategies to mitigate identified harms or risks for each intended use.

Use only verified, structured data sources as UCDP

Think about: Human oversight (as NGOs) before using outputs.

System

System code

Review the code for reliability.

We examined the source code during the development process and after the implementation

Think about: We could call external developers from the development team to review it

Oversight

Automatic oversight

Continuously monitor metrics and utilize guardrails or rollbacks to ensure the system's output stays within a desired range.

We personally tested the system's output and checked that the output was what we wanted

Think about: We could verify the outputs without human supervision

Actions to take

Team

Team formation

Ensure team diversity.

The developmentteam members comefrom different cultural and geographic backgrounds, enriching the understanding of varied contexts

Think about: We could change our academic backgrounds; currently, we are all computer engineers

Team

Team training

Train team members on ethical values and regulations.

The team follows documented ethical guidelines and best practices tailored to the project's goals. Someone from the team attended a data ethics course

Think about: Interaction with ethics committees

Actions taken

Uses

Identification of uses

Work with relevant parties to identify intended uses.

Decision support for applying AI solutions to post-war problems raising awareness of AI applications by suggesting potential risks and potential mitigations for those risks.

Harms

Harm identification

Identify potential harms and risks associated with the intended uses.

There is a risk that outputs generated contain inaccuracies especially if the data are biased or incomplete.
ProMPT may process sensitive conflict-related data.
Recommendations could be used wrongly to justify unetichal behaviors

System

System information

Document all system components, including the AI models, to enable reproducibility and scrutiny.

We have fully described the ProMPT pipeline, including inputs, output formats, and provided examples and code; the LLM models and the reasons for choosing these models; and the databases, which include version control, update procedures, and taxonomy alignment.

System

System evaluation

Report evaluation metrics for various groups based on factors such as age, gender, and ethnicity.

We measured the helpful impact of ProMPT's outputs on a group of people, ranging from very low (1) to very high (5).

Actions taken

System

System information

Document all system components, including the AI models, to enable reproducibility and scrutiny.

We have fully described the ProMPT pipeline, including inputs, output formats, and provided examples and code; the LLM models and the reasons for choosing these models; and the databases, which include version control, update procedures, and taxonomy alignment.

System

System evaluation

Report evaluation metrics for various groups based on factors such as age, gender, and ethnicity.

We measured the helpful impact of ProMPT's outputs on a group of people, ranging from very low (1) to very high (5).

System

System interpretability

Provide mechanisms for interpretable outputs and auditing.

Outputs from each prompt build logically on the previous one, creating a documented chain of reasoning. Impact assessment cards and intermediate outputs reference specific entries from the conflict event and AI use case databases. Comprehensive documentation explains the prompt logic, data sources, and output formats, helping users understand the system's results.

System

System sustainability

Provide an environmental assessment of the system.

ProMPT primarily relies on pre-trained large language models (LLMs) rather than training models from scratch. ProMPT does not perform any model training itself. ProMPT's deployment involves making API calls to the LLM for inference. The total GPU usage corresponds to the aggregate compute time consumed during these calls. Data is stored and processed using lightweight JSON and CSV formats

Actions taken

Data

Dataset information

Compare the quality, representativeness, and fit of training and testing datasets with the intended uses.

The datasets are structured and tagged with a taxonomy aligned to ProMPT's objectives enabling dynamic filtering based on the specific conflict scenario selected by the user. This alignment ensures that the training and testing data directly support generating contextually relevant prompts and impact assessments. For the representativeness the post-conflict dataset is curated to cover a wide range of conflict scenarios. This ensures that the data reflects the complexity of real-world post-conflict environments

Data

Dataset quality

Identify any measurement errors in input data and their associated assumptions.

A measurement error is over-represented or under-represented data or incomplete data. Applying a predefined taxonomy helps standardize input data interpretation, minimizing ambiguity and classification errors.

Data

Dataset protection

Protect sensitive variables in training/testing datasets.

Data Minimization: only variables essential for analysis are included; sensitive variables that are not necessary are excluded to reduce exposure risk.

Oversight

Human oversight

Ensure human control over the system.

The code, datasets and models can be used by anyone. Anyone can take the code and modify it or change the data and model.

Three most difficult cards

- **Automatic oversight**: this card is difficult to answer because we do not have a method to verify that the answer given by the LLM is sensible without the intervention of a man
- **Harm identification**: there is no absolute certainty, and potential issues may emerge over time; therefore, this response may be incomplete. It is essential to continue testing the product to ensure its reliability and robustness
- **System sustainability**: it's hard to estimate how much we consume with API calls; it varies a lot depending on which AI model is used.