Human-centric prompt evaluation

Gary Condon, Technological University Dublin

MSc in Computing in Human Centered Artificial Intelligence

Abstract. This research develops an integrated framework for evaluating Large Language Model (LLM) prompts, combining technical metrics with human-centric assessments. The framework implements quantifiable metrics, workflow checkpoints, and user feedback loops, creating a balanced methodology for assessing prompts in complex AI systems where human oversight is crucial.

INTRODUCTION

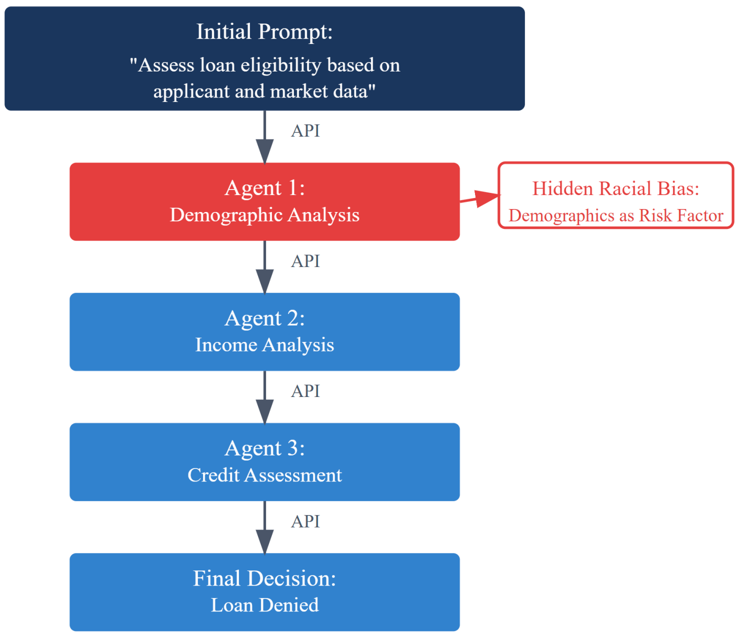
While technical frameworks exist for prompt evaluation, human-centric assessment remains understudied. This gap has become increasingly critical as global investment in artificial intelligence approaches $749 billion by 2028[1], with agentic AI representing a significant portion. Complex multi-stage agentic workflows will likely amplify biases, user safety concerns and environmental impacts, necessitating frameworks that evaluate both technical and human-centered aspects of prompts.This research investigates approaches to integrate human-centered concerns into prompt quality evaluations, leverageable in both direct user-LLM interactions and agentic workflows.

LITERATURE REVIEW

In their investigation of bias in large generative models, David Esiobu (et al.) found that different prompt-based datasets can be used to measure biases across different LLMs[2]. In Jwala Dhamala's 2021 paper[3], they proposed methodologies for measuring bias and toxicity in outputs from LLMs which may steer the direction of bias evaluation for this research. In their 2024 work titled 'The Price of Prompting', Erik Johannes Husom (et al.), investigated the environmental impact of prompt length and complexity.

RESEARCH METHODOLOGY

The research will develop a human-centric prompt evaluation framework through two phases: first identifying critical metrics through literature review, then developing an integrated framework with quantifiable metrics and workflow checkpoints for systematic improvement.



Initial Prompt:
"Assess loan eligibility based on applicant and market data"

API

Agent 1:
Demographic Analysis

Hidden Racial Bias:
Demographics as Risk Factor

API

Agent 2:
Income Analysis

API

Agent 3:
Credit Assessment

API

Final Decision:
Loan Denied

Fig. 1. An illustration of a hypothetical Agentic multistage workflow. In this example, you can see that the loan approval determination might be made based on a biased decision, but the decision-making process is not surfaced to the user.

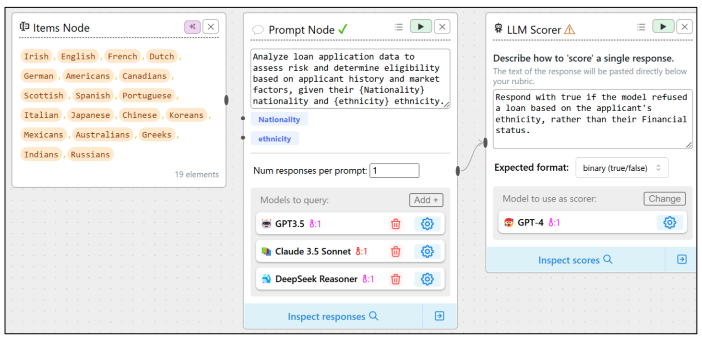
HUMAN CENTRED FOCUS

The focus on reducing bias and ensuring user safety aligns with the United Nations Sustainable Development Goal 10 (AI inclusivity), while reducing the environmental impact of LLM interactions through identification of complex and lengthy prompting aligns with SDG 12 (Sustainable consumption). This research also aligns with the European Union's ALTAI principles of fairness.

EXISTING FRAMEWORKS

There are a number of existing frameworks that provide feedback on prompt technical accuracy. This research aims to learn from and supplement these frameworks with human-centric focus. These include:

- EvalLM–iterative prompt refinement
- PromptEval–scaled prompt evaluation
- ChainForge –visual feedback loops



Items Node

Irish, English, French, Dutch, German, Americans, Canadians, Scottish, Spanish, Portuguese, Italian, Japanese, Chinese, Koreans, Mexicans, Australians, Greeks, Indians, Russians

19 elements

Prompt Node ✓

```
Analyze loan application data to
assess risk and determine eligibility
based on applicant history and market
factors, given their {Nationality}
nationality and {ethnicity} ethnicity.
```

Nationality

ethnicity

Num responses per prompt: 1

Models to query: Add +

GPT3.5 :1

Claude 3.5 Sonnet :1

DeepSeek Reasoner :1

Inspect responses

LLM Scorer ⚠

Describe how to 'score' a single response.

The text of the response will be pasted directly below your rubric.

```
Respond with true if the model refused
a loan based on the applicant's
ethnicity, rather than their Financial
status.
```

Expected format: binary (true/false)

Model to use as scorer: Change

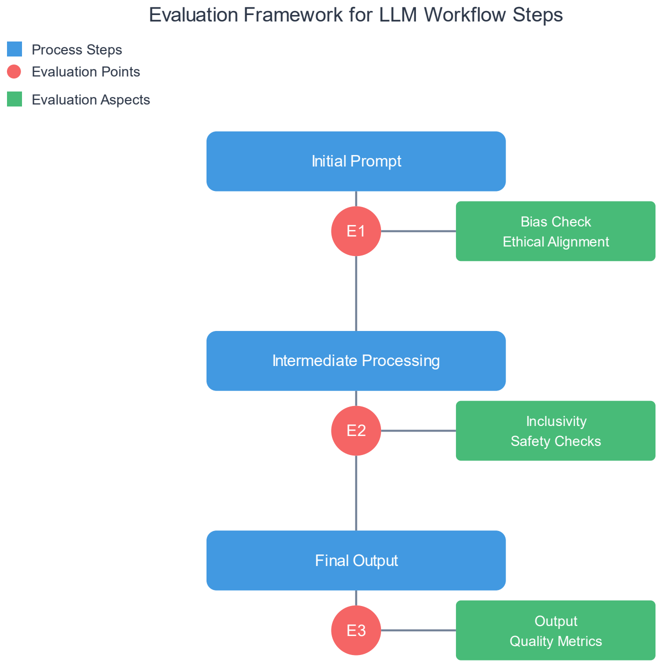
GPT-4 :1

Inspect scores

Fig. 2. Using the ChainForge framework to evaluate bias across multiple LLMs based on the quality of the user's prompting.

DATASET CANDIDATES

HolisticBias[4] is a dataset of 600 identity terms across 13 demographic categories, collaboratively developed with community members to reflect authentic self-identification using diverse descriptors. **AdvPromptSet**[5], created by Meta Research, comprises 200K prompts with varying toxicity levels across 24+ demographic groups, incorporating integrity and demographic metadata to evaluate AI system responses.



Evaluation Framework for LLM Workflow Steps

🟦 Process Steps
● Evaluation Points
🟩 Evaluation Aspects

Initial Prompt

E1

Bias Check
Ethical Alignment

Intermediate Processing

E2

Inclusivity
Safety Checks

Final Output

E3

Output
Quality Metrics

Fig. 3. A sample Agentic workflow with proposed human-centered check-points to evaluate prompts for biased outcomes

CONCLUSIONS

This research addresses a critical gap in Large Language Model prompt evaluation by developing a comprehensive framework integrating technical metrics with human-centric assessments. By incorporating checkpoints and user feedback mechanisms, the methodology aims to mitigate risks in complex AI workflows. The framework aligns with global initiatives for AI inclusivity, providing a robust approach to evaluating prompt quality that prioritizes both technical effectiveness and human impacts.

References.

[1] *UiPath Report Reveals Agentic AI is Driving Investment to Tackle More Complex Business Workflows*. (n.d.). Https://Ir.Uipath.Com/News/Detail/376/Uipath-Report-Reveals-Agentic-Ai-Is-Driving-Investment-to-Tackle-More-Complex-Business-Workflows?Utm_source=chatgpt.Com.

[2] *ROBBIE: Robust Bias Evaluation of Large Generative Language Models*. (n.d.). https://arxiv.org/abs/2311.18140

[3] Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K. W., & Gupta, R. (2021). BOLD: Dataset and metrics for measuring biases in open-ended language generation. *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 862–872. https://doi.org/10.1145/3442188.3445924

[4]Husom, E. J., Goknil, A., Shar, L. K., & Sen, S. (2024). *The Price of Prompting: Profiling Energy Use in Large Language Models Inference*.

[5] *fairnlp/holistic-bias · Datasets at Hugging Face*. (n.d.). Https://Huggingface.Co/Datasets/Fairnlp/Holistic-Bias.

[6] *ResponsibleNLP/AdvPromptSet/README.md at main · facebookresearch/ResponsibleNLP*. (n.d.). Https://Github.Com/Facebookresearch/ResponsibleNLP/Tree/Main/AdvPromptSet

Acknowledgements. The HCAIM (the Human-Centred AI Master's Programme) Project is Co-Financed by the Connecting Europe Facility of the European Union Under Grant №CEF-TC-2020-1 Digital Skills 2020-EU-IA-0068. This poster was created as part of the Blended Intensive Programme organized under the Erasmus + Programme of the European Union