Outline for Form 4 – New Work Item Proposal:

Information Technology – Artificial Intelligence – Guidance for Output Data Quality of Generative AI Application

1. Introduction

AI Applications that are developed using generative AI techniques, models and frameworks are referred to as Generative AI applications. Such applications can create new content or repurpose existing content in the form of text, images, video, etc. This can be processed or incorporated into the output of the application for a specific use-case. The technology enables AI developers to create new AI applications easily. Many of these are powered by large AI models (also now known as foundation models) such as Large Language Models (LLMs), Generative adversarial networks (GANs), Stable Diffusion, Variational Auto Encoders (VAEs), etc.

Generative AI applications differ from traditional AI applications, particularly in terms of the "AI Generated Output". Quality of AI generated output is crucial for adoption of generative AI across use cases.

To ensure the generated data is fit for purpose, a macro-level guidance is necessary to help, understand and interpret generated output for all stakeholders, irrespective of their technical expertise. This helps wider adoption of generative AI across use cases. Since generative AI represents a specialized subset of AI, a use case-based analysis is essential to develop this guidance effectively.

This standard provides guidance for all relevant stakeholders to assess the quality of data generated in the context of development, use and regulation of generative AI applications.

2. Scope

Provide guidance to measure quality of output data generated by generative AI application.

3. Terms and definitions

In this section, relevant generative AI concepts and terms will be taken from existing standards [1-3] to lay the foundation of attributes of a generative AI Application. Additional ones, in the context of generative AI, will be proposed and defined during the development of the standard.

4. Key considerations

The formulation of guidance for output data quality is based on the following: (1) Macro level perspectives of application provider, end user and regulator and its impact on the decision making and scalability of generative AI application; (2) Criteria for assessment.

The criteria for assessment includes metrics and evaluation (including human evaluations).

Stakeholders Involved:

All the stakeholders involved in generative AI application along with their roles and responsibilities will be captured in this clause. Any new stakeholder not already covered in the existing standard [1] to be referred here in line with the supporting use case.

Guidance towards establishing generative AI output data quality:

This clause includes considerations towards establishing generative AI output data quality based on the use cases analysis [4-9]. For example, for the use case involving contract redlining [8], we see that the legal community is leveraging Generative AI to automatically update the contract summary by analyzing original and redlined documents. Generative AI is also being used to generate a consolidated document that provides a summary of all changes made throughout the contract's lifecycle, resulting in a comprehensive Master Contract.

Poor output data quality in this case can lead to inaccurate amendments, missed key clauses, compliance risks, ambiguity, incomplete revisions, reduced efficiency, legal liability, and data privacy concerns. Strict guidelines in ensuring high-quality output is crucial for mitigating these risks and promoting wider adoption of Generative AI with legal community.

The guidance towards measuring generative AI output data quality will include following key considerations.

Explainability: Understanding the "why" behind a model's decision. It sheds light on the reasoning process used to arrive at generated output providing insights into how factors influence output, allowing users to trust the outputs.

Safety: This involves considerations towards providing robust performance and safe use of generative AI applications while minimizing the negative impact, through multiple checks like Profanity, toxicity, etc.

Fairness and Bias: Identification of Bias at different stages and ensuring fairness as the subjective practice of using generative AI without favoritism or discrimination.

Privacy: Generative AI output should protect Personal Identifiable Information (PII) information and include detection mechanisms for data leakage and privacy preserving techniques such as anonymization.

Context awareness and sensitivity: This involves assessment of the generated text with respect to incorporation of contextual information and adaptation to changes in the input context. Context-sensitive outputs demonstrate a deeper understanding of the task and context.

Adversarial Robustness: Measures the model's resilience against adversarial inputs or attacks intended to manipulate or exploit its behaviour. Adversarial robust outputs maintain quality and coherence even when facing malicious inputs.

Diversity: Reflects the variety and novelty of generated outputs. Higher diversity indicates a broader range of ideas and expressions, enhancing creativity and preventing repetition.

Contextual Hallucination: Contextual hallucination highlights the limitations of LLMs in truly understanding context and generating responses that reflect genuine comprehension or knowledge.

Perplexity Score: Perplexity is a measure on how well the model predicts the next word or character based on the context provided by the previous words or characters. The lower the perplexity score, the better the model’s ability to predict the next word accurately.

Uncertainty: Uncertainty refers to the LLM's lack of confidence in the correctness or accuracy of its generated text. When an LLM is presented with a prompt, it can generate multiple possible continuations or completions. The uncertainty associated with each completion reflects the LLM's internal assessment of how likely each continuation is to be the most appropriate one.

Token Importance: To calculates the importance of individual tokens in a text prompt by comparing the original embedding with an embedding where a single token is removed.

Coherence: This refers to collective quality of all the sentences including contextual relevance, cohesive structure, consistent tone and style and logically organized.

Consistency: This refers to the factual alignment between the summary and the summarized source. A factually consistent summary contains only statements that are entailed by the source document. GPT model is used as an evaluator, and this is a prompt-based evaluation method.

Relevance: This refers to the selection of important content from the source. The summary should include only important information from the source document. Annotators were instructed to penalize summaries which contained redundancies and excess information. GPT model is used as an evaluator, and this is a prompt-based evaluation method.

Fluency: This refers to the quality of the summary in terms of grammar, spelling, punctuation, word choice, and sentence structure. GPT model is used as an evaluator, and this is a prompt-based evaluation method.

Faithfulness**:** This primarily assesses consistency, contextual understanding, and adherence to any given constraints of the model. This can be better understood through tone and style, relevance and coherence of response with the context

Factuality: This focuses on the verifiability and validation of evidence, response plausibility, error detection, and alignment with established knowledge graphs. This can be achieved by relying on latest/updated information and credible or factual databases/sources, if there is any bias identified, reference to any existing knowledge graphs.

Note: *With respect to the above considerations, this section will evaluate alignment of generative AI with existing standards for traditional AI. Accordingly, the existing developments to be referred and new considerations specific to generative AI to be highlighted.*

References

[1] ISO/IEC 22989: 2022 – Information Technology – Artificial Intelligence Concepts and Terminology.

[2] ISO/IEC 5338:2023 – Information technology — Artificial intelligence — AI system life cycle processes

[3] ISO/IEC 5339:2024– Information technology — Artificial intelligence — Guidance for AI applications

[4] Use case on generative AI, ISO-IEC JTC 1-SC 42-WG 4_N1648

[5] Use case on generative AI, ISO-IEC JTC 1-SC 42-WG 4_N1649

[6] Use case on generative AI, ISO-IEC JTC 1-SC 42-WG 4_N1650

[7] Use case on generative AI, ISO-IEC JTC 1-SC 42-WG 4_N1651

[8] Use case on generative AI, ISO-IEC JTC 1-SC 42-WG 4_N1652

[9] IN expert contribution - Generative AI - Data Quality, ISO-IEC JTC 1-SC 42-WG 4_N1647