

[illegible]

Hung	Walking a Tightrope - Evaluating Large L	2023	b	Safety	general	High risk domains often involve sensitive. Safety	Safety is defined as the degree of other	SafetyRisk control	Two threads: Detourally and Safe. We consider Detourally & Safet	Not applicable	Safety of generated language, evals	2
Kano	Integrated NLP Evaluation System for PI	2009		Decomposit	general	To understand the key characteristics of I	Not reported	fully automatic	Integrated system that can: Not reported	Not applicable	Evaluation of NLP annotations using	2
Kano	Decomposit: Evaluating Generative Models	2020		general	general	Not reported	Not reported	fully automatic	Not reported	Not applicable	Decomposit: Generalized open-source tool	2
Khabazi	GENE: Toward Reproducible and Standard	2020		Human evaluation/crowdsources	general	While often assumed a good standard, off	Standardized system	general	GENE, Amazon Mechanical Turk, and	Not applicable	GENE, Amazon Mechanical Turk, and	2
Kao	Automated Assessment of Fidelity and	2024	a	Fidelity	general	As Large Language Models (LLMs) become	Fidelity pertains to how accurately	fully automatic	Fidelity is based on faithfulness. Two primary components of fidelity	Not applicable	Assessment of fidelity based on me	2
Kao	Li: A Large-Scale Study on the Reliability	2023	a	Interpretability	general	Most existing work mainly focuses on the	Interpretability assesses how accurately	LLM	Interpretability assesses how accurately	Not applicable	Assessment of these domains of me	2
Liu	G-EVAL: Large-scale evaluation using GPT-4	2024		general	general	"Conventional approaches to assess the	The quality of texts generated by the LLM	general	Worked, sentence databases, "Only need the	Not applicable	The authors conducted a benchmark	2
Macharia	Improving Text Generation Through Intero	2024		coherence	general	The main purpose of this paper is to assess	Whether the texts generated by the LLM	general	Whether the texts generated by the LLM	Not applicable	Whether the texts generated by the LLM	2
Mehrotra	Evaluating Large Language Models as Agents	2024		general	general	Rather than relying on external tools, the	Whether the texts generated by the LLM	general	Whether the texts generated by the LLM	Not applicable	Whether the texts generated by the LLM	2
Min	FACTSCORE: Fine-grained Atomic Eval	2023	a	Human FACTSC	general	Evaluating the factuality of long form text	Factual precision—whether each p	human	Generated text is broken into. Comparing FACTSCORE using Atomic FACTSC	Not applicable	Human evaluation of the factuality	1
Min	FACTSCORE: Fine-grained Atomic Eval	2023	a	Human FACTSC	general	Evaluating the factuality of long form text	Factual precision—whether each p	human	Generated text is broken into. Comparing FACTSCORE using Atomic FACTSC	Not applicable	Human evaluation of the factuality	1
Mohammedi	Active Evaluation Framework	2022		Human evaluation	general	Given a system, a human approaches to	Whether the texts generated by the LLM	general	Whether the texts generated by the LLM	Not applicable	Whether the texts generated by the LLM	2
Morido	Assessing image and text generation w	2021	a	Quality	general	Objective and interpretable metrics to	Whether the texts generated by the LLM	general	Whether the texts generated by the LLM	Not applicable	Whether the texts generated by the LLM	2
Morido	Assessing image and text generation w	2021	a	Quality	general	Objective and interpretable metrics to	Whether the texts generated by the LLM	general	Whether the texts generated by the LLM	Not applicable	Whether the texts generated by the LLM	2
Morido	Assessing image and text generation w	2021	a	Quality	general	Objective and interpretable metrics to	Whether the texts generated by the LLM	general	Whether the texts generated by the LLM	Not applicable	Whether the texts generated by the LLM	2
Morido	Assessing image and text generation w	2021	a	Quality	general	Objective and interpretable metrics to	Whether the texts generated by the LLM	general	Whether the texts generated by the LLM	Not applicable	Whether the texts generated by the LLM	2
Morido	Assessing image and text generation w	2021	a	Quality	general	Objective and interpretable metrics to	Whether the texts generated by the LLM	general	Whether the texts generated by the LLM	Not applicable	Whether the texts generated by the LLM	2
Morido	Assessing image and text generation w	2021	a	Quality	general	Objective and interpretable metrics to	Whether the texts generated by the LLM	general	Whether the texts generated by the LLM	Not applicable	Whether the texts generated by the LLM	2
Morido	Assessing image and text generation w	2021	a	Quality	general	Objective and interpretable metrics to	Whether the texts generated by the LLM	general	Whether the texts generated by the LLM	Not applicable	Whether the texts generated by the LLM	2
Morido	Assessing image and text generation w	2021	a	Quality	general	Objective and interpretable metrics to	Whether the texts generated by the LLM	general	Whether the texts generated by the LLM	Not applicable	Whether the texts generated by the LLM	2
Morido	Assessing image and text generation w	2021	a	Quality	general	Objective and interpretable metrics to	Whether the texts generated by the LLM	general	Whether the texts generated by the LLM	Not applicable	Whether the texts generated by the LLM	2
Morido	Assessing image and text generation w	2021	a	Quality	general	Objective and interpretable metrics to	Whether the texts generated by the LLM	general	Whether the texts generated by the LLM	Not applicable	Whether the texts generated by the LLM	2
Morido	Assessing image and text generation w	2021	a	Quality	general	Objective and interpretable metrics to	Whether the texts generated by the LLM	general	Whether the texts generated by the LLM	Not applicable	Whether the texts generated by the LLM	2
Morido	Assessing image and text generation w	2021	a	Quality	general	Objective and interpretable metrics to	Whether the texts generated by the LLM	general	Whether the texts generated by the LLM	Not applicable	Whether the texts generated by the LLM	2
Morido	Assessing image and text generation w	2021	a	Quality	general	Objective and interpretable metrics to	Whether the texts generated by the LLM	general	Whether the texts generated by the LLM	Not applicable	Whether the texts generated by the LLM	2
Morido	Assessing image and text generation w	2021	a	Quality	general	Objective and interpretable metrics to	Whether the texts generated by the LLM	general	Whether the texts generated by the LLM	Not applicable	Whether the texts generated by the LLM	2
Morido	Assessing image and text generation w	2021	a	Quality	general	Objective and interpretable metrics to	Whether the texts generated by the LLM	general	Whether the texts generated by the LLM	Not applicable	Whether the texts generated by the LLM	2
Morido	Assessing image and text generation w	2021	a	Quality	general	Objective and interpretable metrics to	Whether the texts generated by the LLM	general	Whether the texts generated by the LLM	Not applicable	Whether the texts generated by the LLM	2
Morido	Assessing image and text generation w	2021	a	Quality	general	Objective and interpretable metrics to	Whether the texts generated by the LLM	general	Whether the texts generated by the LLM	Not applicable	Whether the texts generated by the LLM	2
Morido	Assessing image and text generation w	2021	a	Quality	general	Objective and interpretable metrics to	Whether the texts generated by the LLM	general	Whether the texts generated by the LLM	Not applicable	Whether the texts generated by the LLM	2
Morido	Assessing image and text generation w	2021	a	Quality	general	Objective and interpretable metrics to	Whether the texts generated by the LLM	general	Whether the texts generated by the LLM	Not applicable	Whether the texts generated by the LLM	2
Morido	Assessing image and text generation w	2021	a	Quality	general	Objective and interpretable metrics to	Whether the texts generated by the LLM	general	Whether the texts generated by the LLM	Not applicable	Whether the texts generated by the LLM	2
Morido	Assessing image and text generation w	2021	a	Quality	general	Objective and interpretable metrics to	Whether the texts generated by the LLM	general	Whether the texts generated by the LLM	Not applicable	Whether the texts generated by the LLM	2
Morido	Assessing image and text generation w	2021	a	Quality	general	Objective and interpretable metrics to	Whether the texts generated by the LLM	general	Whether the texts generated by the LLM	Not applicable	Whether the texts generated by the LLM	2
Morido	Assessing image and text generation w	2021	a									

[illegible]

[illegible]

[illegible]

