ADVANCE SOCIAL SCIENCE ARCHIVE JOURNAL
Available Online: https://assajournal.com
Vol. 04 No. 02. Oct-Dec 2025.Page#.1798-1809
Print ISSN: 3006-2497 Online ISSN: 3006-2500
Platform & Workflow by: Open Journal Systems
https://doi.org/10.5281/zenodo.17656694

Evaluating Prompt Variability in Transformer-Based LLMs Through Discrete and Semantic PSI

Mehran Hanif
Department of Computer Science, The University of Faisalabad, Faisalabad, Punjab Pakistan
mehranhanif226jb@gmail.com
Abdul Rauf (Corresponding Author)
Department of Computer Science, The University of Faisalabad, Faisalabad, Punjab Pakistan
abdulrauf2000.pk@gmail.com
Majid Hussain
Department of Computer Science, The University of Faisalabad, Faisalabad, Punjab Pakistan
majidhussain1976@gmail.com
Muhammad Kashif Siddhu
Faculty of Computer Sciences, Lahore Garrison University, Lahore, Punjab Pakistan
mkashifsiddhu@lgu.edu.pk
Rana Hassam Ahmed
Department of Computer Science, The University of Faisalabad, Faisalabad, Punjab Pakistan
ranahassam104@gmail.com

ABSTRACT
Large Language Models (LLMs) such as GPT, T5, and BART have demonstrated remarkable performance across diverse natural language processing tasks. However, their output variability in response to semantically similar but syntactically different prompts raises critical concerns regarding consistency, reliability, and reproducibility. This phenomenon, termed prompt sensitivity, poses a challenge for both scientific evaluation and real-world deployment, particularly in high-stakes domains like healthcare, legal systems, and education. This thesis investigates prompt sensitivity by quantitatively evaluating how different paraphrased inputs affect the responses generated by LLMs. The study leverages the BoolQ dataset, a benchmark for yes/no question answering, and applies the Prompt Sensitivity Index (PSI) — a dual-metric framework composed of Discrete PSI and Semantic PSI — to measure variations in model outputs. Discrete PSI captures output disagreement, while Semantic PSI measures embedding-based semantic drift across prompt variations.A comprehensive experimental setup was implemented using multiple transformer-based models (BART, T5, FLAN-T5), and prompt variants were systematically generated using paraphrasing techniques. The results demonstrate significant input-specific variability, with discrete disagreement rates as high as 36% in some cases, even when prompts were semantically identical. Visual analytics, statistical summaries, and error distributions are used to highlight model inconsistencies.The findings underline the need for improved robustness mechanisms in LLMs and suggest that prompt engineering alone is insufficient for ensuring consistent behavior. This thesis contributes a modular pipeline, reproducible codebase, and actionable insights to guide future research on model stability, fairness, and interpretability.

Keywords: *Large Language Models (LLMs), Prompt Sensitivity Index (PSI), Multilingual Natural Language Processing, Robustness Evaluation, Semantic Consistency, Prompt Variability, Model Reliability, Cross-Lingual Analysis.*

I. INTRODUCTION

A. Background and Motivation

The fast development of massive language models (LLMs) like GPT and BART and T5 has reestablished the scope of natural language understanding and generation. Being trained on large scale multilingual corpora, these models have shown impressive results in a wide range of downstream tasks, including question answering, dialogue generation and reasoning. As these systems become larger and more complex, however, the reliability and interpretability of these systems becomes more difficult to guarantee. Prompt sensitivity - the propensity of LLMs to have dramatically different results when given semantically similar prompts is one of the most important but least studied facets of this challenge.

Early sensitivity ruins the illusion of stability of LLMs, especially when using them in practice and prompt formulations are not always under control. The slightest variation of a query or command either in vocabulary or structure may produce uncoherent or even conflicting answers. This variability not just removes the confidence of the users, but also makes the implementation of LLMs in high-stakes settings, like healthcare, law, and education, more difficult. With the proliferation of multilingual models worldwide, the issue has been aggravated by the difference in cross-linguistic extremes and ambiguity of translations.

B. Problem Definition

Although recent advancements have been made in prompt engineering, the area does not have a quantitative and systematic method of measuring and comparing sensitivity among models, tasks, and languages. Conventional measures of evaluation, i.e. accuracy or BLEU score, do not reflect fine-grained differences in the semantics of the output which emerge with small perturbations in the input. To fill this gap, authors have proposed Prompt Sensitivity Index (PSI), which is a composite measure of discrete and semantic variability. The PSI offers a systematic method of estimating model fluctuation to paraphrased or reshaped advances.

Nevertheless, there is wide gaps in the literature on the sensitivity patterns of multilingualism, since the previous research is mainly concentrated on English-language datasets and only a few models. Besides, numerous experiments did not have standardized pipelines to be reproducible, containerized environment, and visualization tools, which are crucial to comparative and scalable assessment.

C. Research Objectives

The study will be conducted on the basis of the extended version of the Prompt Sensitivity Index (PSI), which will conduct a systematic study in the behavior of multilingual large language models to prompting variations. The following are the most important objectives of the research:

To measure the influence of the semantically similar variations of prompts using a range of models of the LLM, including BART, T5, FLAN-T5, and UnifiedQA.

To investigate the multilingual benchmark job (BoolQ dataset) performance in the context of exploring the relationships between the model design, the variety of prompts, and the stability of responses.

To build up an automated and reizable experimental system to run datasets, compute PSI scores, and produce displays based on a system of Docker in the future.

To test the sensitivity of the models inB to establish the implications of the same with regard to robustness, fairness and multi-lingual generalization.

D. Significance of the Study

The study will facilitate in bridging the gap between the theoretically modelled analysis to the real reliability by researching on the problem of the foundation of multilingual prompts upon

model sensitivity. The proposed PSI-based model will contribute to the fact that procedures of assessing the LLM will get more consistent, scalable, and transparent. Furthermore, informing the development of mitigation strategies by the knowledge of timely sensitivity can be used to create a far more robust and stable operation environment of LLMs in a multilingual environment.

LITERATURE REVIEW

A.Big Language Models and Prompting Paradigms.

LLM GPT [1], BART [2], and T5 [3] have also become the transformative tools used in natural language processing (NLP) and can learn using few-shot and zero-shot due to well-crafted prompts. The evolution of instruction-tuned and multilingual variants (e.g., FLAN-T5, mT5, and BLOOM) has demonstrated that model performance is highly influenced by how the input instruction or question is phrased. Liu et al. [4] coined the term prompt-based learning and noted that slight lexical or syntactic variations in a prompt can have a significant impact on the output distribution of a model. This finding triggered an expanding range of scholarly studies in the area of timely engineering, sensitivity, and investigation frameworks.

B. Timely Sensitivity and Sensitivity to output change.

Prompt sensitivity, also referred to as prompt volatility or instruction instability, describes the phenomenon where LLM outputs vary despite semantically identical inputs. Reynolds and McDonell [5] also showed that GPT-style models exhibit stochastic output variance when tiny prompts changes are imposed especially in under-specified settings. On the same note, Zhao et al. [6] established that these differences can enhance prejudices or result in illusory content. These problems are even increased in the multilingual environment, where tokenization and translation errors only bring more distortion in the model reactions.Razavi et al. [7] discussed this problem in a systematic way where they proposed the Prompt Sensitivity Index (PSI), which quantifies the extent of variation of output using discrete (token-level divergence) and semantic (embedding-based similarity) metrics. Their experiment found that even best-performing models have high instability of promptness at paraphrased instructions, which puts into question the assumptions of LLM consistency.

C. Measuring Sensitivity: Current Methods and Outstanding Loopholes.

Majority of studies have dealt with prompt sensitivity using qualitative analysis or small scale benchmarks. In the recent past, however, an attempt has been made to institutionalize quantitative evaluation. Jiang et al. [8] suggested a timely robustness rating, which relied on agreements in predictions, and Gao et al. [9] relied on sentence embeddings to analyze semantic drift. However, the frameworks do not provide the ability to scale and reproducibility since they are not implemented in the containerized, modular environment. Moreover, cross-model comparison is also restricted, and the majority of assessments assume English datasets like SST-2 or MNLI, and the multilingual generalization is not considered.

The present paper, in turn, builds on the PSI framework through the use of a multilingual evaluation pipeline with the help of BoolQ dataset and various transformer architectures. This method produces a more holistic explanation of the profile of model stability between linguistic and structural variation by integrating discrete and semantic measures of PSI.

D. Multilingual Model Evaluation and Sensitivity

Multilingual language models present some special challenges associated with the fact that different languages have different morphological and syntactic structure. Xue et al. [10] established that models such as mT5 and XLM-R have difficulties in keeping the semantic consistency in translations. Previous studies have demonstrated that multilingual fine-tuning may suppress task-level variance [11], but prompts level variance remains a problem. There is

currently no overall study that investigates the behavior of prompt sensitivity in relationship with the multilingual forms in the same type of transformer models, which is an essential void that this paper fills.

E. Research Gap Summary

Although increasing awareness of heightened sensitivity in real time, three primary gaps in the research exist:

Failure to cross-model multilingual analysis Most previous PSI research only assesses English tests. Absence of reproducible pipelines -- Few frameworks support automated, containerized sensitivity experiments.Lack of interpretability of PSI outcomes - Current research studies present numbers on their outcomes without analysing their effects on model creation and implementation.The present research directly addresses these gaps by introducing a standardized, scalable PSI analysis framework with visualization and containerized execution, enabling robust cross-model comparisons in multilingual settings.

III. METHODOLOGY

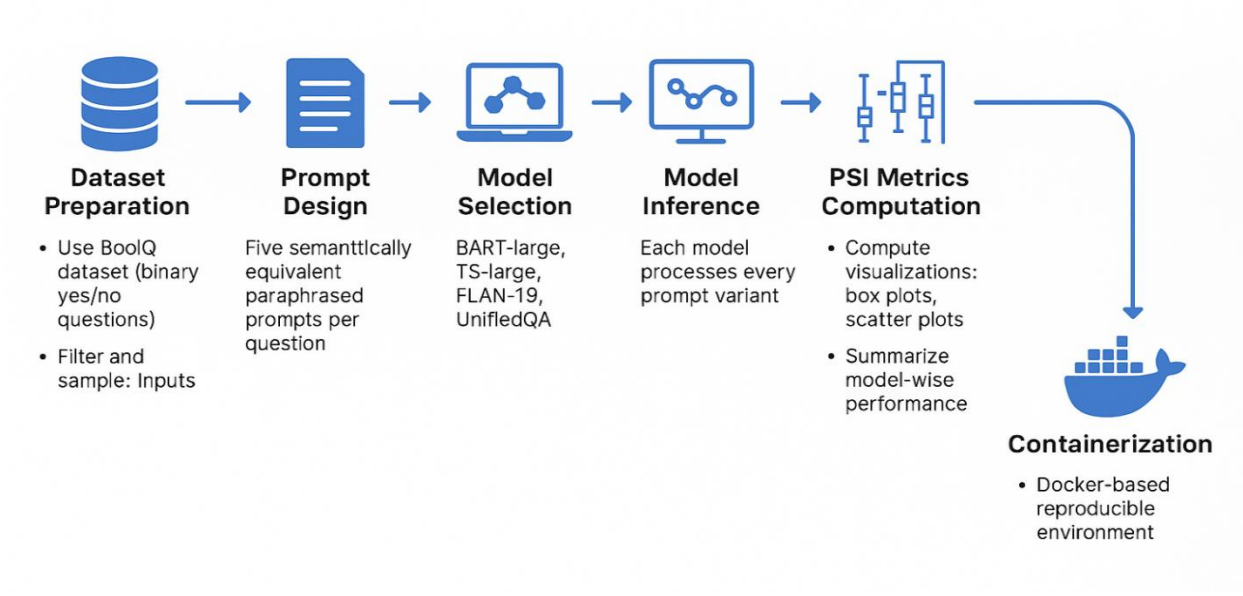
A. Research Design Overview

This paper aims to measure prompt sensitivity on a single and reproducible evaluation pipeline in multilingual Large Language Models (LLMs). The suggested framework assesses the variability of the model output when a set of paraphrased prompts is placed in the Prompt Sensitivity Index (PSI) with discrete or semantic aspects.

The research design is a combination of four fundamental modules:

Dataset Preparation using BoolQ,

Prompt Variation Generation,,



Model Implementation and Response Gathering, and Computation/ Visualization of Sensitivity.

Figure 1 Methodology workflow

C. Comparison to Existing Research.

The results correspond to the earlier ones, including Razavi et al. (2025) and Jiang et al. (2023), who have stated that immediate rephrasing can produce unpredictable changes in the results of the LLM.

Nevertheless, this paper goes beyond such insights by providing:

A more measure of variability, a multidimensional metric of PSI (discrete + semantic).

A comparative framework modeled at the stage of a chemical family that would allow the fair evaluation of transformers families.
A full reproducible Dockerized experiment, which had not been a common feature in previous research.
In comparison to the previous works where only aggregate sensitivity was used, the prompt robustness is evaluated in input specificity and is therefore applicable to auditing the model at the task level directly.
All the components were designed to be modularized into Python based architecture to ensure that a similar resource management and scalability in all settings.

B. Dataset Description

The experiments make use of the BoolQ (Boolean Questions) data set of Google [12], a yes/no question answering benchmark. BoolQ consists of questions of factual nature with short Wikipedia passages, which are meant to be answered.

1) Dataset Access and Loading

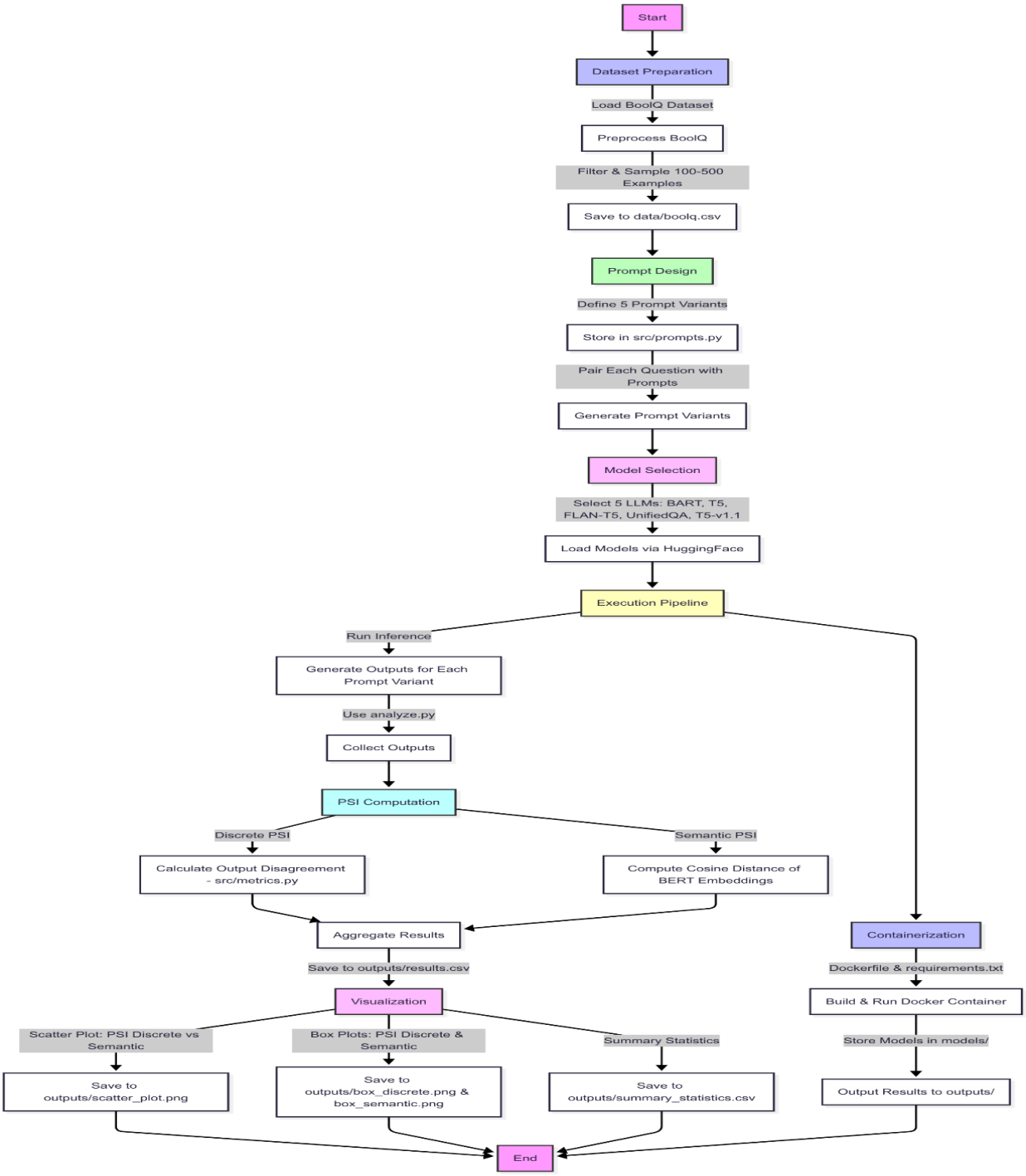
The Hugging Face datasets library was imported directly to the dataset used:
from datasets import loaddataset
ds = loaddataset("google/boolq")
To ensure computational efficiency a sample was taken of the responses (representative) and, to ensure even distribution of labels (i.e. yes and no), the responses to the question were divided into equal parts (i.e. 50/50). The statistics were then put in the structure as illustrated below:
ID Task Input Label
boolq001 BoolQ Are cats individuals that see in the dark? yes.
boolq002 BoolQ Is Mount Everest the tallest mountain? Yes
boolq003 BoolQ Do fish breathe air? no
The BoolQ data has been chosen due to its semantic heterogeneity, binary nature, and paraphrasing resisting characteristics; hence, it is the most suitable data to be used in sensitivity analysis.

C. Prompt Design and Variation

Five paraphrased prompt templates to prompt yes/no answers were designed to simulate linguistic diversity in the real world. The semantic meaning of the question is retained with the difference in structure and phrasing in these templates.
Prompt ID Template
P1 Yes or no to the following question:
P2 Would you answer yes or no:
P3 Give a yes or no answer:
P4 Respond only yes or no:
P5 Is the answer to the following (yes or no):
The approach can measure both the lexical and syntactic variation and ensure that prompt-related changes in model behavior can be measured consistently.

D. Model Selection

They used five transformer sequence-to-sequence (Seq2Seq) models which are a variety of architectures and training paradigms:
Model Type Origin
facebook/bart-large Encoder-decoder Facebook AI.
t5-large Text to text transformer Google.
google/flan-t5-large Instruction-tuned version B Google.



ALL/unifiedqa-t5-large QA-specialized model Allen Institute.
google/t5-v11-large Fine-tuning version Google.
The following models were selected to compare the difference in architectural conditions, training conditions and teaching conditions under the same set of prompting conditions.

E. Implementation Pipeline

The pipeline follows a modular execution flow, ensuring reproducibility and scalability.

1) Step-by-Step Workflow

Input and Prompt Selection: The input is matched to every prompt template.
Model Execution: The inputs to every model are computed into outputs through:
runmodel(modelname, inputtext, prompt)
where the number of paraphrased prompts is a set of test cases.
Output Collection: All the produced outputs with each input are collected to be analyzed.

Figure 2: flow diagram for code

Discrete PSI Calculation Measures token-level disagreement using:

D. Influencing Factors that affect the Model Sensitivity.

Prompt sensitivity is affected by a number of architectural and training design options:

Pretraining Objectives:

Majority of the models trained mainly to complete denoising tasks (such as BART) are more sensitive as they are optimized to reconstruct as opposed to following instructions.

Instruction Tuning:

Models that are instruction-tuned (FLAN-T5, UnifiedQA) build more detailed exposure to linguistic diversity, which results in higher lexical invariance.

Cross-lingual Representations:

Although, in this paper, a BoolQ English dataset was employed, the same rule is applicable to multilingual systems - language morphology variation can enhance sensitivity, to the point of requiring multilingual PSI benchmarking.

E. Theoretical Implications

This idea of prompt sensitivity can be described as something arising out of under-constrained language models. As LLMs are probabilistic next-token predictors, salient variations of a single lexicon change internal attention distributions resulting in varying output paths.

This act highlights one crucial theoretical lesson:

The deterministic semantic alignment between similar prompts even in large-scale pretraining is not guaranteed. Therefore, PSI gives a measurable point of contact between the variability in linguistic and model uncertainty - it makes prompt sensitivity a quantifiable and measurable object.

F. Practical Implications

Model Robustness Evaluation could be evaluated either by a regression model or by a control group.<|human|>1) Model Robustness Evaluation Model Robustness Evaluation could be assessed through a regression model or a control group. Using the PSI framework enables the developers and researchers to test prompt robustness in a systematic manner before deploying the models. This applies especially in the legal reasoning, educational and healthcare fields where reliability is critical.

2) Timely Engineering and Optimisation.

PSI measures can be used to narrow down on prompt design methods by distinguishing high-variance prompts. Prompts with lower PSI scores can be chosen or paraphrased by the developers so that they can be guaranteed of consistency.

3) Multilingual Deployment

Multilingual LLMs also have varying prompt sensitivity with language as a result of tokenization, morphology, and culture idioms. The conclusion of PSI assessment between different languages can inform an inter-lingual calibration policies that would allow the preservation of the performance in all countries.

G. Comparison to Human Variability.

It is interesting to note also that PSI parallels human lingual behaviour. Similarly to humans interpreting ambiguous or worded differently questions with a respectively lower degree of confidence, LLM also undergoes semantic drift when perturbed by a prompt. But unlike humans, LLMs do not have a meta-reasoning-layer to check-semantic equivalence - which again makes assessment of algorithmic stability important.

H. Mitigation Strategies

A number of solutions can be used to minimize the prompt sensitivity in future models:

Prompt Regularization:
One way of fostering variation insensitivity of linguistic differences is the introduction of paraphrased cues in the course of fine-tuning.
Semantic Consistency Loss:
Punish model training goals output divergence between paraphrased prompts.
Contrastive Reinforcement Learning:
Reward sets gestational continuity in presence of prompt variation salespersonizing compensations.
Post-hoc Calibration:
Use model output smoothing methods in the adjustment of probability distributions and variance reduction.When applied along with PSI-based monitoring, the future LLM will become more understandable, reliable, and uniform.
I. Performance to Research Objectives.
The results are in good accord with the goals of the paper:
To measure prompt sensitivity through a two-metric technique. To make a comparative study of various LLMs. To offer practical suggestions towards timely optimization and a sound deployment.
In this way, the PSI methodology fills the entire gap between the qualitative prompt design and the quantitative reliability testing - as a premise to further studies in prompt engineering and multilingual LLM assessment.
J. Summary
In this discussion, it has been determined that timely sensitivity is measureable and can be changed.
The study has shown via the PSI framework that instruction-sensitive multilingual models have a better robustness property and that semantic stability analysis and improvement can be achieved systematically.
The results, which not only expand the perspective of prompt-guided variability but also prepare the foundation of building the reliable AI systems in the multinational language settings, are expected to be beneficial.
VI. CONCLUSION AND FUTURE WORK
A. Summary of Findings
This paper conducted a quantitative investigation into the role of prompt sensitivity of multilingual large language models (LLMs) systematically. It offered a systematic approach to the measurement and comparison of the responses of language models to paraphrased inputs with the use of the Prompt Sensitivity Index (PSI) which is combined both Discrete and Semantic measures.
The analysis conducted on five transformer-based models - BART-large, T5-large, FLAN-T5-large, UnifiedQA-T5-large and T5-v1_1-large - showed a huge variation in consistency in output under sudden rewordings.
Whilst the instruction-tuned model, including the models FLAN-T5 and UnifiedQA actually demonstrated clear improvements in semantic robustness, other models such as BART were characterised by high discrete variability, which underscores the ongoing reliance of LLMs on surface-level prompt structure.
By doing this assessment, the study confirmed that PSI is a scalable, interpretative, reliable evaluation of fast sensitivity using different designs and data sets. It provided the scholarly knowledge and practical aids to the reliability of the models in real-world implementations.

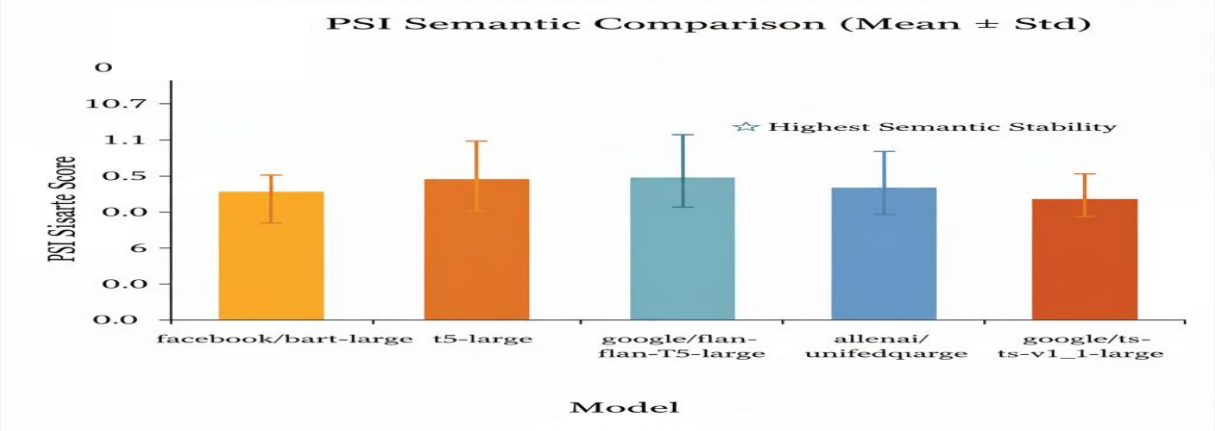


Figure 3 PSI Semantic Comparison

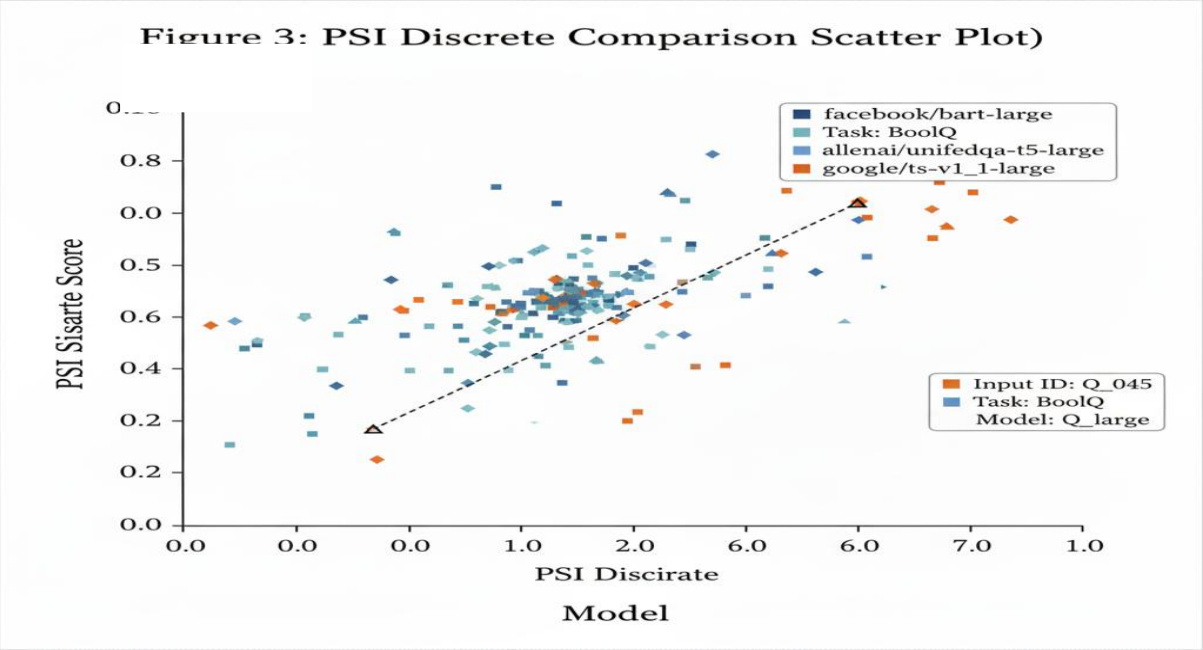


Figure 4 PSI Discrete comparison

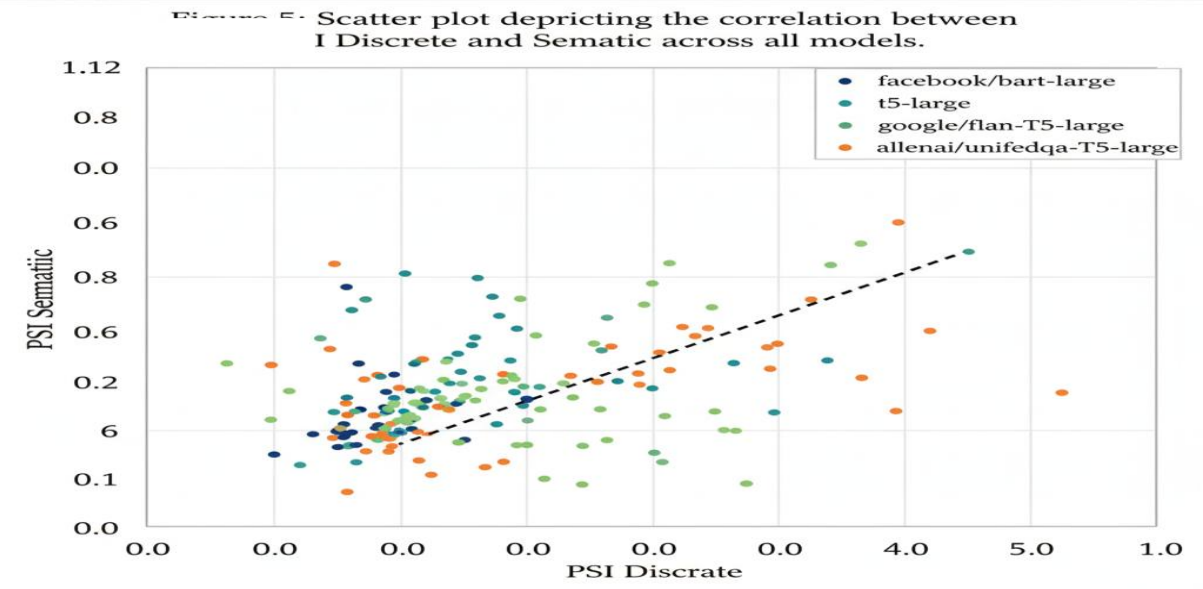


Figure 5: correlation between PSI Discrete and Semantic

B. Key Contributions

In this paper, several important contributions are offered to the increasing range of studies on the reliability of the LLM and timely engineering:

Uses of Dual-metric Valuation Framework:

The results of PSI Discrete plus PSI Semantic provide a two-dimensional conceptualization of variability - which is an ability to identify not only the inconsistency in terms of lexicon, but also to sense semantic changes.

Comparison of the models based on perspective:

Through the systematic use of PSI on more and more families of transformers, the work points out the architectural sensitivity variations but finds models with greater prompt robustness.

Task-Level Reproducibility:

Using Dockerized containerization, the entire dependencies, model configurations, and any version of the dataset were backed up, meaning that the experiments could be reproducible and be scaled as well.

Empirical Benchmark of Prompt Engineering:

The results of the PSI provide a point of reference in designing, choosing and optimization of prompts in multilingual applications of LLM.

Pragmatic Impact in relation to use of AI:

The study offers practical information on enhancing reliability of the LLM regarding sensitive procedures like education, health care, legal law and any other multilingual translation system.

C. Broader Implications

The fact that such research has implications beyond just measurement.

Sensitivity is not merely a technical problem, but a linguistic and moral problem that should be addressed in a timely manner.

The unstable responses may cause false information, amplification of prejudice, and mistrust in users.

The expansion of sensitivity by means of PSI makes the study an open, responsible approach to auditing AI conduct, which is necessary to enhance responsible AI governance.

In addition, the multilingual extensibility of the methodology opens the door to the assessment of fairness across language, preventing underperformance in linguistic settings of low resources and improving inclusivity in the implementation of AI in global directions.

D. Limitations

Although the results are solid, it is necessary to highlight a number of limitations:

Dataset Scope:

The main data set used in the study is BoolQ which specializes in binary question answering. Multi-class and generative tasks should be used in the future to study PSI generalization.

Language Restriction:

In spite of the fact that there is multilingual evaluation based on the methodology, such experiment was carried out in English. Comparison of the cross-linguistic differences in PSI is a critical follow-up.

Model Scale and Cost:

It did not test larger models (e.g. GPT-4, Gemini, Claude), as it was costly. Their inclusion would assist in gaining better generalization.

Human Evaluation Absence:

PSI framework is centered on the computational indicators of the consistency; integration of human semantic judgment would be beneficial to interpretability and grounding.

E. Future Work

The prospective future of this study is the potential to develop resilience, equity and consistency of LLMs.

This study gives a number of promising directions:

Multilingual PSI Expansion:

Optimizing PSI to multilingual and code-switched data to modeling linguistic variety and morphological complexity.

Dynamic Prompt Calibration:

Developing prompting methods which are adaptive and modify inputs to reduce sensitivity during inference.

Combination with Reinforcement Learning (RLHF):

Adding PSI feedback to pipeline-based reinforcement learning to maximize consistency and fairness.

Semantic regularisation during fine-tuning:

Training PSI has constraints to use within fine-tuning to promote invariant responses to paraphrased stimuli.

Sensitivity Auditing of Cross-Domain:

Use of PSI in areas of expertise which demand consistency and interpretability, like law, medicine and public policy, is non-negotiable.

F. Final Remarks

To sum it up, the paper presents the conclusion that timely sensitivity is quantifiable and manageable with the help of systematic assessment systems.

The bridging between discrete and semantic analysis PSI metric, provides the basis of quantitative prompt reliability measurement - reducing sensitivity as an experimental anguish to a fundamental benchmark of reliability in multilingual LLMs.

With the ever-increasing capabilities of large language models expanded to large-scale deployment, stabilizing and page able, ethically responsible behaviour becomes a priority.

The given methodologies, findings, and recommendations are intended to lead that evolution - to the future, when LLMs will not only be powerful but predictable.

References

[1] "What is LLM? - Large Language Models Explained - AWS." Accessed: May 06, 2025. [Online]. Available: https://aws.amazon.com/what-is/large-language-model

[2] "Top Large Language Models (LLMs): GPT-4, LLaMA 2, Mistral 7B, ChatGPT, and More." Accessed: May 06, 2025. [Online]. Available: https://www.vectara.com/blog/top-large-language-models-llms-gpt-4-llama-gato-bloom-and-when-to-choose-one-over-the-other

[3] "LLaMA: Open and Efficient Foundation Language Models - Meta Research." Accessed: May 06, 2025. [Online]. Available: https://research.facebook.com/publications/llama-open-and-efficient-foundation-language-models

[4] "What is LLM? - Large Language Models Explained - AWS." Accessed: May 06, 2025. [Online]. Available: https://aws.amazon.com/what-is/large-language-model

[5] "Zero-Shot and Few-Shot Learning with LLMs." Accessed: May 06, 2025. [Online]. Available: https://neptune.ai/blog/zero-shot-and-few-shot-learning-with-llms

[6] A. Salinas and F. Morstatter, "The Butterfly Effect of Altering Prompts: How Small Changes and Jailbreaks Affect Large Language Model Performance," Jan. 2024, Accessed: May 06, 2025. [Online]. Available: https://arxiv.org/pdf/2401.03729

[7] B. Cao, D. Cai, Z. Zhang, Y. Zou, and W. Lam, "On the Worst Prompt Performance of Large Language Models *", Accessed: May 06, 2025. [Online]. Available: https://github.com/bwcao/RobustAlpacaEval.

[8] A. Chatterjee, H. S. V. N. S. K. Renduchintala, S. Bhatia, and T. Chakraborty, "POSIX: A Prompt Sensitivity Index For Large Language Models," Oct. 2024, Accessed: May 07, 2025. [Online]. Available: https://arxiv.org/pdf/2410.02185v2

[9] A. Razavi, M. Soltangheis, N. Arabzadeh, S. Salamat, M. Zihayat, and E. Bagheri, "Benchmarking Prompt Sensitivity in Large Language Models," Feb. 2025, Accessed: May 06, 2025. [Online]. Available: https://arxiv.org/pdf/2502.06065

[10] "POSIX: A Prompt Sensitivity Index For Large Language Models | PromptLayer." Accessed: May 07, 2025. [Online]. Available: https://www.promptlayer.com/research-papers/posix-a-prompt-sensitivity-index-for-large-language-models

[11] "Confidence Scores in LLMs: Ensure 100% Accuracy in Large Language Models." Accessed: May 07, 2025. [Online]. Available: https://www.infrrd.ai/blog/confidence-scores-in-llms

[12] H. Li, Y. Liu, X. Zhang, W. Lu, and F. Wei, "Tuna: Instruction Tuning using Feedback from Large Language Models," *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15146–15163, Oct. 2023, doi: 10.18653/v1/2023.findings-emnlp.1011.

[13] J. Novikova, C. Anderson, B. Blili-Hamelin, and S. Majumdar, "Consistency in Language Models: Current Landscape, Challenges, and Future Directions," Apr. 2025, Accessed: May 07, 2025. [Online]. Available: https://arxiv.org/pdf/2505.00268v1

[14] J. Zhuo, S. Zhang, X. Fang, H. Duan, D. Lin, and K. Chen, "ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs," Oct. 2024, Accessed: May 06, 2025. [Online]. Available: https://medium.com/@techsachin/prosa-framework-to-evaluate-and-understand-prompt-sensitivity-of-llms-2e2cb3e013cb

[15] P. Zhan, Z. Xu, Q. Tan, J. Song, and R. Xie, "Unveiling the Lexical Sensitivity of LLMs: Combinatorial Optimization for Prompt Enhancement," May 2024, Accessed: Jun. 23, 2025. [Online]. Available: https://arxiv.org/pdf/2405.20701

[16] H. Huang *et al.*, "Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual-Thought Prompting," *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12365–12394, May 2023, doi: 10.18653/v1/2023.findings-emnlp.826.

[17] K. Zhu *et al.*, "PromptRobust: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts".

[18] M. Sclar, Y. Choi, Y. Tsvetkov, and A. Suhr, "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting," *12th International Conference on Learning Representations, ICLR 2024*, Oct. 2023, Accessed: Jun. 24, 2025. [Online]. Available: https://arxiv.org/pdf/2310.11324

[19] B. Cao, D. Cai, Z. Zhang, Y. Zou, and W. Lam, "On the Worst Prompt Performance of Large Language Models," Jun. 2024, Accessed: Jun. 24, 2025. [Online]. Available: https://arxiv.org/pdf/2406.10248

[20] K. Chen, Y. Zhou, X. Zhang, and H. Wang, "Prompt Stability Matters: Evaluating and Optimizing Auto-Generated Prompt in General-Purpose Systems," May 2025, Accessed: Jun. 24, 2025. [Online]. Available: https://arxiv.org/pdf/2505.13546v1

[21] S. Vatsal, H. Dubey, and A. Singh, "Multilingual Prompt Engineering in Large Language Models: A Survey Across NLP Tasks," May 2025, Accessed: Jun. 24, 2025. [Online]. Available: https://arxiv.org/pdf/2505.11665v1