# EVALUATING CONCURRENT ROBUSTNESS OF LANGUAGE MODELS ACROSS DIVERSE CHALLENGE SETS

Vatsal Gupta[1†], Pranshu Pandya[1†], Tushar Kataria[2], Vivek Gupta[3], Dan Roth[4]

[1] Indian Institute of Technology Guwahati, [2] University of Utah, [3] Arizona State University, [4] University of Pennsylvania





EMNLP 2024

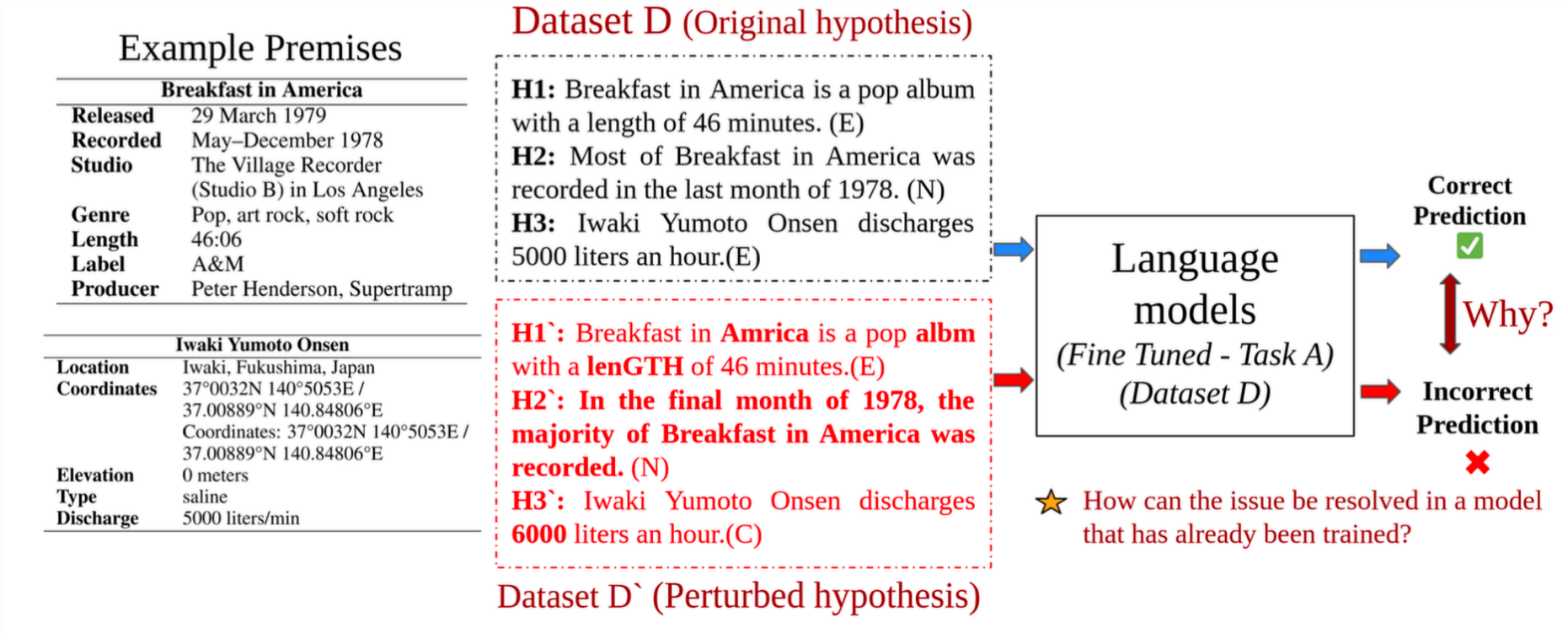
# INTRODUCTION

- **Reliability:** As LMs continue to be integrated into daily life, ensuring their reliability is crucial.
- **Sensitivity to Minor Perturbations:** LMs often display unexpected behavior when faced with slight changes in input.
- **Extending to Concurrent Perturbations:** Extensive analysis has been performed on single-set inoculation. However, models trained on one type of pertrubation fail to be generalizable to other types.
- **Methodology for Analysis:** We propose the **Multi-Set inoculation framework**, generalizable across different NLP tasks to analyze and improve concurrent robustness.


EMNLP 2024

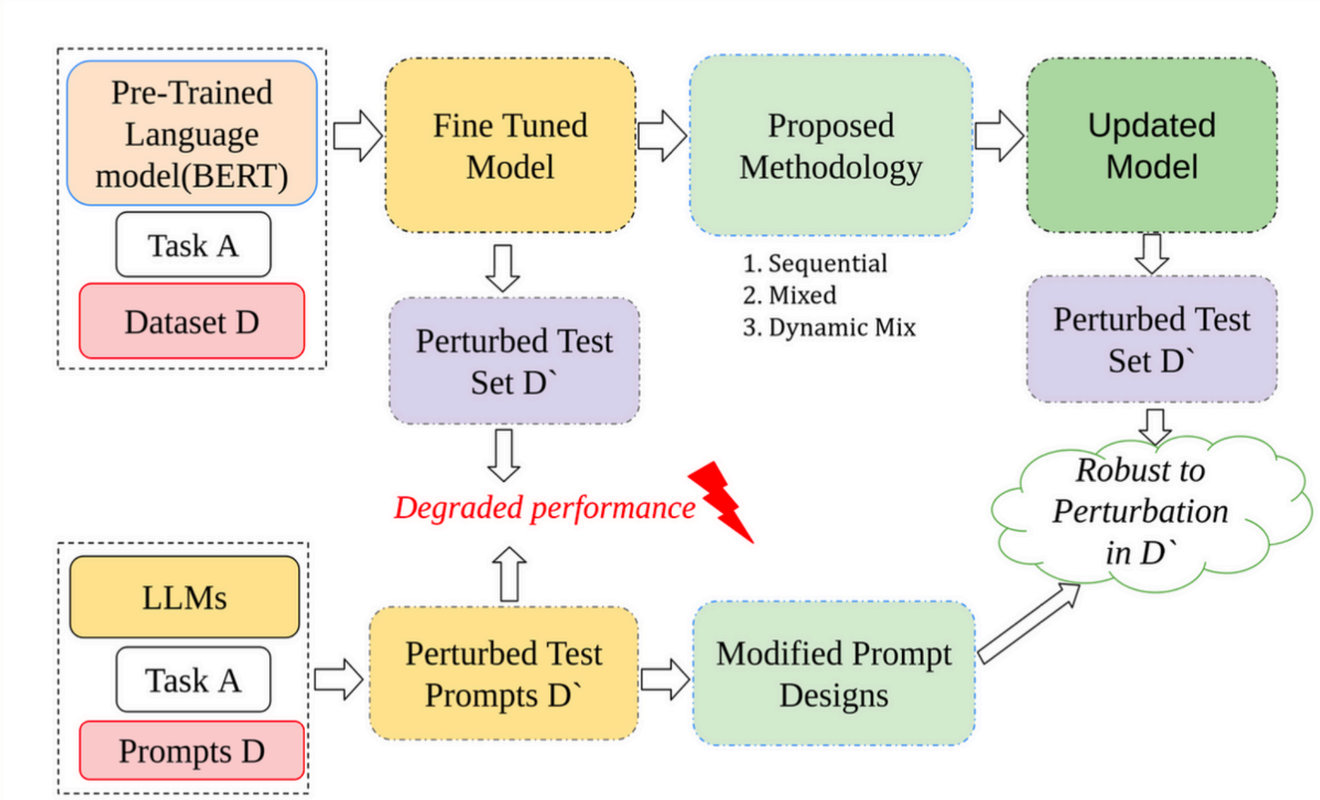
# PROBLEM STATEMENT

- How can a **Multi-Set Inouclation Framework** be proposed i.e. a model be made robust to multiple perturbations and how can the concurrent robustenss problem be addressed and analysed ?
- What **possible strategies** can work and how do they compare against one another in different settings?
- How can we propose **cost-effective strategies for PLMs and LLMs which are generalizable to different NLP tasks ?**




EMNLP 2024

# METHODOLOGY




**For PLMs**

- Initial Fine-Tuning Start with a pre-trained language model, fine-tuned on original unperturbed dataset for a specific task.
- Challenge Sets Create perturbed test sets by applying different input perturbations
- Inoculation Strategies:
  - **Sequential Training**: Fine-tune models sequentially one by one on each challenge set.
  - **Mixed Training**: Fine-tune on a combined set with examples from all perturbations.
  - **Dynamic Mix**: Adjust sample ratios based on perturbation difficulty, which allows for hard sample selection.

EMNLP 2024

# METHODOLOGY

**For LLMs**

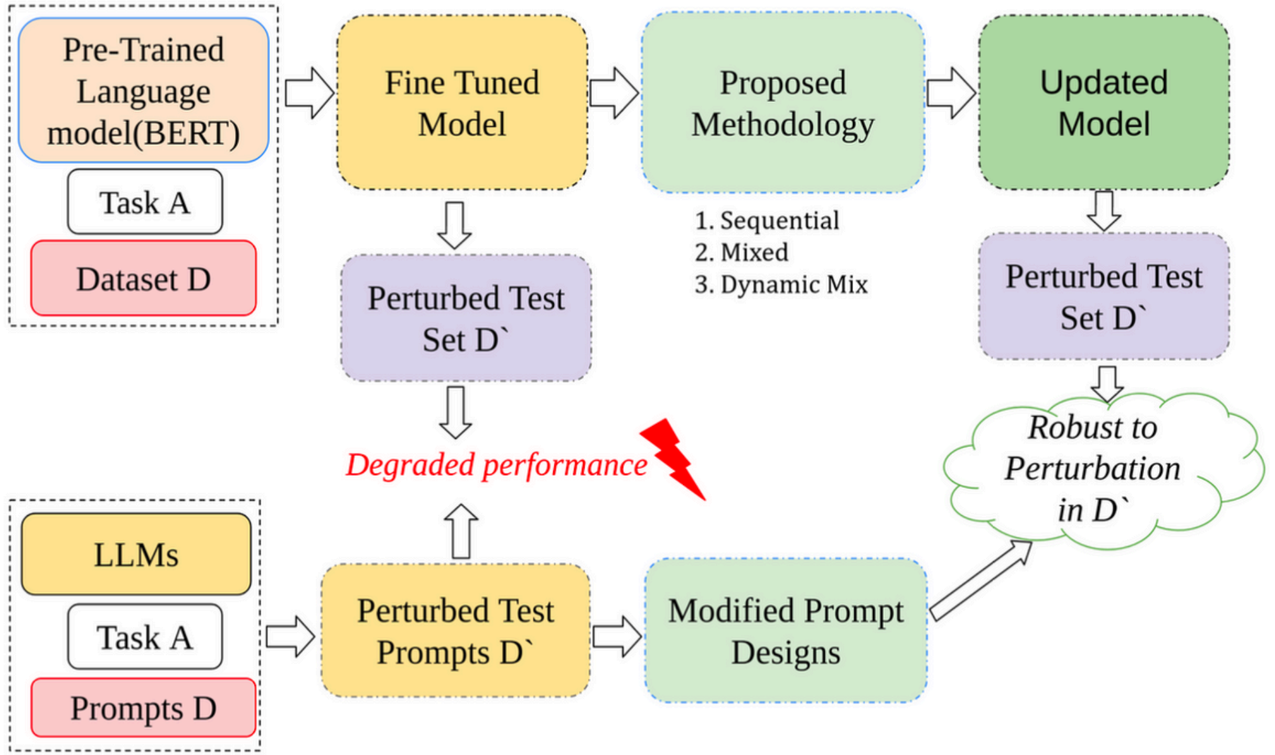
- Few-Shot Chain of Thought Prompting**:** Task exemplars with reasoning chains improve resilience by enhancing contextual understanding.

- Perturbation-Aware Prompting

  - **Single Exemplars Multiple Prompts (SEMP):** Tailored prompt with description and exemplars for each perturbation type.

  - **Multiple Exemplars Single Prompt (MESP):**
    - Have detailed exemplars and descriptions covering all perturbation types and make model elicit reasoning.

    - **MESP(MPI)**: Focus more on detailed descriptions
    - **MESP(MPE)**: Focus more on exemplars along with description of mistake and correction.



# CASE-STUDY ON TABULAR NLI

| Case Closed | |
|---|---|
| **Written** | Takahiro Arai |
| **Publish** | Shogakukan |
| **Eng. Publish** | SG Shogakukan Asia |
| **Demographic** | Shonen |
| **Magazine** | Weekly Shonen Sunday |
| **Orig. Run** | May 9, 2018 - present |
| **Volumes** | 2 (List of volumes) |

$H_1$: Takahiro Arai wrote 'Case Closed' comic series. (E)
$H_1'$: Takahiro Arai wotte 'Case Closed' comci series. (E)
$H_2$: 'Case Closed' is a long-term comic series.(E)
$H_2'$:'Case Closed' isn't a long-term comic series.(C)
$H_3$:'Case Closed' became the anime Detective Conan (N)
$H_3'$:Detective Conan is 'Case Closed' anime version. (N)
$H_4$:'Case Closed' has run over 5 years.(E)
$H_4'$:'Case Closed' has run over 10 years.(C)
$H_5$: Shogakukan Asia published 'Case Closed' (Eng). (E)
$H_5'$:Shogakukan UK published 'Case Closed' (Eng). (C)

- **Tabular NLI task:** We utilize the Tabular-NLI dataset, **INFOTABS** along with 5 perturbations created on it, along with original sets (α1,α2 ,α3) which aren't included in training in any way [1].
  - Character Perturbation: Minor character changes without altering semantics .
  - Numeric Perturbation : Modifies numbers in a sentence.
  - Negation pertrubation : Negates the sentence.
  - Paraphrasing pertrubation:Paraphrases the given sentences.
  - Location Perturbation:Replaces locations in sentences.

- Figure shows examples with Original hypotheses (H) and perturbed hypothesis (H') representing character, negation, paraphrasing, numeric and location perturbations respectively.


Reference:
[1] Abhilash Shankarampeta, Vivek Gupta, and Shuo Zhang. 2022. Enhancing tabular reasoning with pattern exploiting training. In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Asso- ciation for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 706–726, Online only. Association for Computational Linguistics.
[2] Alex Kulesza and Ben Taskar. 2011. k-dpps: Fixed-size determinantal point processes. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), pages 1193–1200.
[3] Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2020. Unsupervised alignment-based iterative ev- idence retrieval for multi-hop question answering. In Proceedings of the 58th Annual Meeting of the Asso- ciation for Computational Linguistics, pages 4514– 4525, Online. Association for Computational Lin guistics.
[4]J. Neeraja, Vivek Gupta, and Vivek Srikumar. 2021. Incorporating external knowledge to enhance tabular reasoning. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2799–2809, Online. Association for Computational Linguistics.


EMNLP 2024

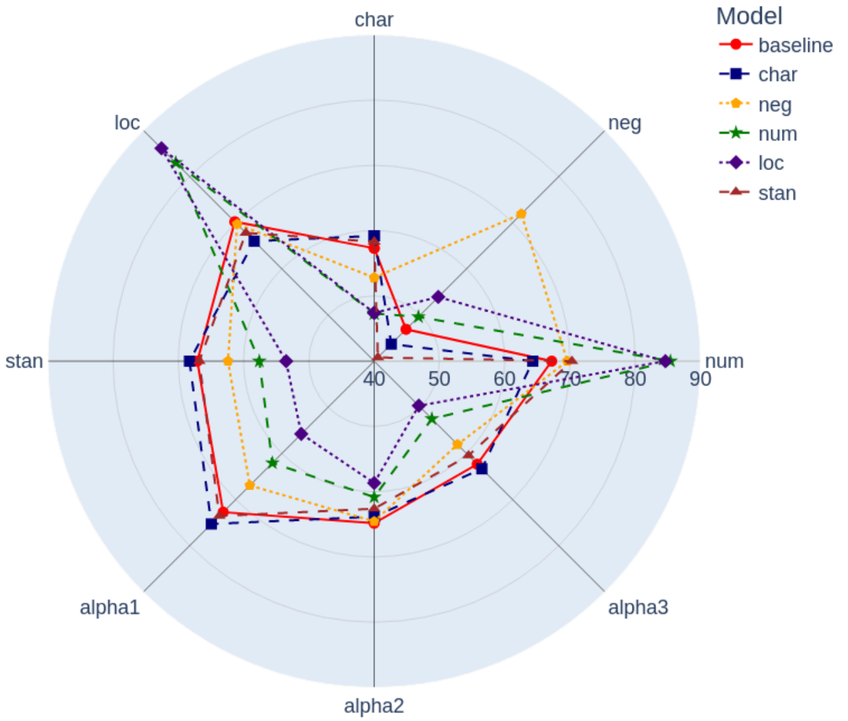
# CASE-STUDY ON TABULAR NLI

- **Sample Selection:** Only pertinent samples post perturbations are selected[2].
- **Tabular Premise Representation:** We employed alignment techniques[3] eliminate distracting rows (DRR) and represent the tabular premise in paragraph form [4].
- **Metric for evaluation :** We label the hypothesis as **Entailment**, **Contradiction** or **Neutral** and the improvement over the concurrent perturbation setting is considered by taking the average of the improved performance over each challenge set from the performance on the untrained model (*Baseline*).
- We analyse performance on both **Pre-Trained Language models (PLMs) and Large Language Models (LLMs)**


Reference:
[1] Abhilash Shankarampeta, Vivek Gupta, and Shuo Zhang. 2022. Enhancing tabular reasoning with pattern exploiting training. In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Asso- ciation for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 706–726, Online only. Association for Computational Linguistics.
[2] Alex Kulesza and Ben Taskar. 2011. k-dpps: Fixed-size determinantal point processes. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), pages 1193–1200.
[3] Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2020. Unsupervised alignment-based iterative ev- idence retrieval for multi-hop question answering. In Proceedings of the 58th Annual Meeting of the Asso- ciation for Computational Linguistics, pages 4514– 4525, Online. Association for Computational Lin guistics.
[4]J. Neeraja, Vivek Gupta, and Vivek Srikumar. 2021. Incorporating external knowledge to enhance tabular reasoning. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2799–2809, Online. Association for Computational Linguistics.


EMNLP 2024

# RESULTS AND ANALYSIS - ESTABLISHING NECESSITY




- **<u>Multi Model Uniset Inoculation</u>**
  - Baseline performance on challenge sets is notably lower than on original sets, emphasizing PLMs' vulnerability to input perturbations.
  - While the fine-tuned model excels against respective perturbations, it struggles with other pertrubations.
- This experiment demonstrates and validates **the need for concurrent robustness** the consequent requirement for coming up with **Multi-Set Inoculation Framework** which includes strategies to address this issue.

EMNLP 2024

# RESULTS AND ANALYSIS - PLMS [PRE-TRAINED LANGUAGE MODELS]

| | | Original Sets | | | Challenge Sets | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $K$/SEQ-Type | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | char | neg | num | loc | stan | $\mu$ |
| | baseline | 72.72 | 64.83 | **62.33** | 57.30 | 46.90 | 67.20 | 70.20 | 67.10 | - |
| SEQ | *COL-ASC* | 61.67 | 60.94 | 50.11 | 48.80 | 54.60 | **85.40** | 85.40 | 56.60 | 4.42 |
| SEQ | *COL-DSC* | **74.67** | 62.72 | 60.44 | **58.90** | 57.30 | 56.10 | 65.30 | **68.00** | -0.62 |
| SEQ | *ROW-ASC* | 55.00 | 58.11 | 47.22 | 46.80 | 50.90 | 84.50 | **85.90** | 51.30 | 2.14 |
| SEQ | *ROW-DSC* | 73.44 | 63.39 | 57.44 | 56.50 | 45.10 | 60.00 | 71.60 | 65.80 | -1.94 |
| MIX | 100 | 70.40 | **65.16** | 59.48 | 56.00 | 58.48 | 78.78 | 78.50 | 66.04 | 5.82 |
| MIX | 200 | 70.42 | 65.06 | 59.21 | 56.86 | 59.50 | 80.94 | 80.36 | 64.68 | 6.73 |
| MIX | 300 | 71.92 | 64.54 | 59.49 | 56.50 | 61.30 | 81.22 | 79.68 | 65.12 | 7.02 |
| MIX | 400 | 72.11 | 64.48 | 59.78 | 56.58 | 63.70 | 81.60 | 80.38 | 64.64 | 7.64 |
| MIX | 500 | 72.62 | 64.34 | 59.20 | 56.98 | **66.06** | 82.02 | 80.52 | 65.64 | **8.50** |
| DYNMIX | 500 | 71.28 | 64.42 | 60.39 | 56.26 | 59.22 | 77.84 | 76.24 | 65.38 | 5.25 |
| DYNMIX | 1000 | 71.07 | 64.72 | 59.60 | 57.04 | 63.24 | 79.94 | 79.06 | 65.50 | 7.22 |
| DYNMIX | 1500 | 72.07 | 64.81 | 59.73 | 56.50 | 65.42 | 80.84 | 79.54 | 65.64 | 7.85 |

Table 3: **Single Model Multi Set Fine tuning Strategies Results:** For SEQ Results , ROBERTA$_{\text{INTA}}$ is Sequential Trained with 500 samples from each P$_j$. Here, COL-ASC: CSNLM, COL-DSC: MLNSC, ROW-ASC: SCNML, ROW-DSC: LMNCS are the sequence types and $\mu$ is the average improvement. For MIX Results, ROBERTA$_{\text{INTA}}$ fine-tuned on $K$ equal samples from different perturbation sets P$_j$. For DYNMIX Results, ROBERTA$_{\text{INTA}}$ fine-tuned on total of $K$ samples taken from P$_j$ in ratios mentioned in the DYNMIX SECTION BELOW.

EMNLP 2024

# RESULTS AND ANALYSIS - PLMS [PRE-TRAINED LANGUAGE MODELS]

| | In-distribution | | | Out-distribution | | Original Test sets | | | |
|---|---|---|---|---|---|---|---|---|---|
| **K** | **neg** | **num** | **loc** | **char** | **stan** | **alpha1** | **alpha2** | **alpha3** | $\mu$ |
| **baseline** | 46.90 | 67.20 | 70.20 | **57.30** | **67.10** | **72.72** | **64.83** | **62.33** | - |
| **100** | 60.4 | 83.2 | 81.4 | 49.6 | 59.6 | 63.6 | 62.8 | 56.1 | 5.10 |
| **200** | 61.9 | 85.6 | 83.0 | 49.2 | 58.0 | 61.3 | 61.9 | 53.0 | 5.79 |
| **300** | 62.1 | 85.8 | 83.2 | 48.8 | 55.7 | 59.4 | 62.3 | 51.9 | 5.39 |
| **400** | 66.3 | 85.1 | 83.5 | 47.5 | 54.3 | 58.4 | 61.5 | 51.1 | 5.61 |
| **500** | **68.0** | **86.0** | **84.1** | 47.8 | 53.9 | 58.0 | 61.2 | 50.1 | **6.23** |

(a) **Fine Tuning on Perturbation Subset (neg, num, loc).** Model fine tuned using MIX strategy using only 3 perturbations. Performance reported on out of distribution perturbation and alpha test sets.

| | In-distribution | | Out-distribution | | | Original Test sets | | | |
|---|---|---|---|---|---|---|---|---|---|
| **K** | **char** | **num** | **neg** | **loc** | **stan** | **alpha1** | **alpha2** | **alpha3** | $\mu$ |
| **baseline** | 57.30 | 67.20 | 46.90 | 70.20 | 67.10 | **72.72** | **64.83** | **62.33** | - |
| **100** | 56.3 | 80.1 | 50.3 | 74.6 | 65.4 | 71.0 | 63.2 | 60.1 | 3.61 |
| **200** | 57.2 | 82.8 | 47.9 | 76.3 | 65.3 | 70.9 | 63.5 | 59.2 | 4.15 |
| **300** | 57.0 | 83.1 | 47.0 | 77.1 | 65.2 | 71.1 | 63.1 | 58.1 | 4.13 |
| **400** | **58.0** | **84.1** | **48.5** | **78.0** | **64.4** | 70.8 | 63.8 | 58.4 | **4.86** |
| **500** | 57.0 | **84.1** | 46.7 | 77.7 | **64.4** | 70.9 | 63.2 | 58.0 | 4.25 |

(b) **Fine Tuning on Perturbation Subset (char, num).** Model fine tuned using MIX strategy using only 2 perturbations. Performance reported on out of distribution perturbation and alpha test sets.

**Ablation Study** Testing Mixed Training on Out of Distribution Perturbation Set

# RESULTS AND ANALYSIS - LLMS [LARGE LANGUAGE MODELS]

Comparing the results of challenge datasets(Q) and their unperturbed version sets(Q') reveals that LLMs similar to PLMs are also sensitive to input data perturbations.

| | | Model | char | neg | num | loc | stan | avg. |
|---|---|---|---|---|---|---|---|---|
| $OP_{ZS}$ | Q' | Flan-t5-XXL | **70.60** | **77.30** | **69.00** | **74.00** | **79.00** | **73.98** |
| | | LLaMA-2-70b | 59.00 | 63.60 | 64.60 | 67.00 | 60.00 | 62.84 |
| | | GPT-3.5 | 68.00 | 69.00 | 68.66 | 71.60 | 70.00 | 69.45 |
| | Q | Flan-t5-XXL | 63.00 | 70.00 | 63.00 | 65.00 | 69.30 | 66.06 |
| | | LLaMA-2-70b | 54.00 | 51.60 | 49.60 | 57.00 | 54.30 | 53.30 |
| | | GPT-3.5 | 51.00 | 53.00 | 62.66 | 61.00 | 60.30 | 57.59 |
| $OP_{CoT}$ | Q' | LLaMA-2-13b | 63.67 | 69.33 | 66.33 | 61.00 | 61.00 | 64.27 |
| | | LLaMA-2-70b | 68.6 | 72.3 | 76.3 | 67.3 | 69.6 | 70.82 |
| | | GPT-3.5 | 68.30 | 76.30 | 68.00 | 73.00 | 75.30 | 72.18 |
| | Q | LLaMA-2-13b | 61.33 | 57.00 | 57.67 | 59.33 | 60.00 | 59.07 |
| | | LLaMA-2-70b | **63.00** | 60.00 | **63.00** | **61.30** | 66.00 | 62.66 |
| | | GPT-3.5 | 63.00 | **69.60** | 59.30 | 61.00 | **68.00** | 64.18 |

Table 4: (a) **Zero Shot ($OP_{ZS}$):** Baseline Accuracies on original and perturbed sets for prompts in zero-shot setting. (b) **Few-shot with CoT ($OP_{CoT}$):** Results using CoT prompting with exemplars sampled from O.

# RESULTS AND ANALYSIS - LLMS

| Pr/ Test | char | neg | num | loc | stan | $Q'$ |
|---|---|---|---|---|---|---|
| baseline | 51.00 | 53.00 | 62.66 | 61.00 | 60.30 | 69.05 |
| char | **67.60** | 65.30 | **66.00** | **69.00** | **67.60** | 68.05 |
| neg | 60.30 | 64.60 | 58.00 | 59.60 | 63.30 | 71.62 |
| num | 62.30 | 66.30 | 61.00 | 60.60 | 64.30 | 70.24 |
| loc | 62.60 | 63.60 | 61.00 | 59.30 | 64.00 | 71.30 |
| stan | 59.00 | **67.60** | 61.30 | 61.00 | 67.30 | **73.76** |

(a) SEMP **Results on GPT-3.5**

| Type | $\pi_j$ | char | neg | num | loc | stan |
|---|---|---|---|---|---|---|
| **BASE** | $Q'_j$ | 59.00 | 63.60 | 64.60 | 67.00 | 60.00 |
| | $Q_j$ | 54.00 | 51.60 | 49.60 | 57.00 | 54.30 |
| **SEMP** | $Q'_j$ | 69.00 | 71.00 | 72.00 | 72.30 | 68.60 |
| | $Q_j$ | 53.00 | 58.00 | 62.00 | 62.00 | 68.30 |

(b) SEMP **Results on LLaMA-2-70b**

Table 5: SEMP **Results**: (a) The last column is the average performance on all sets of Q′ (b) Self-testing on perturbation $\pi_j$ with prompt for $\pi_j$ and test on $Q_j$ and $Q'_j$.

**Single Exemplars Multiple Prompts (SEMP)**

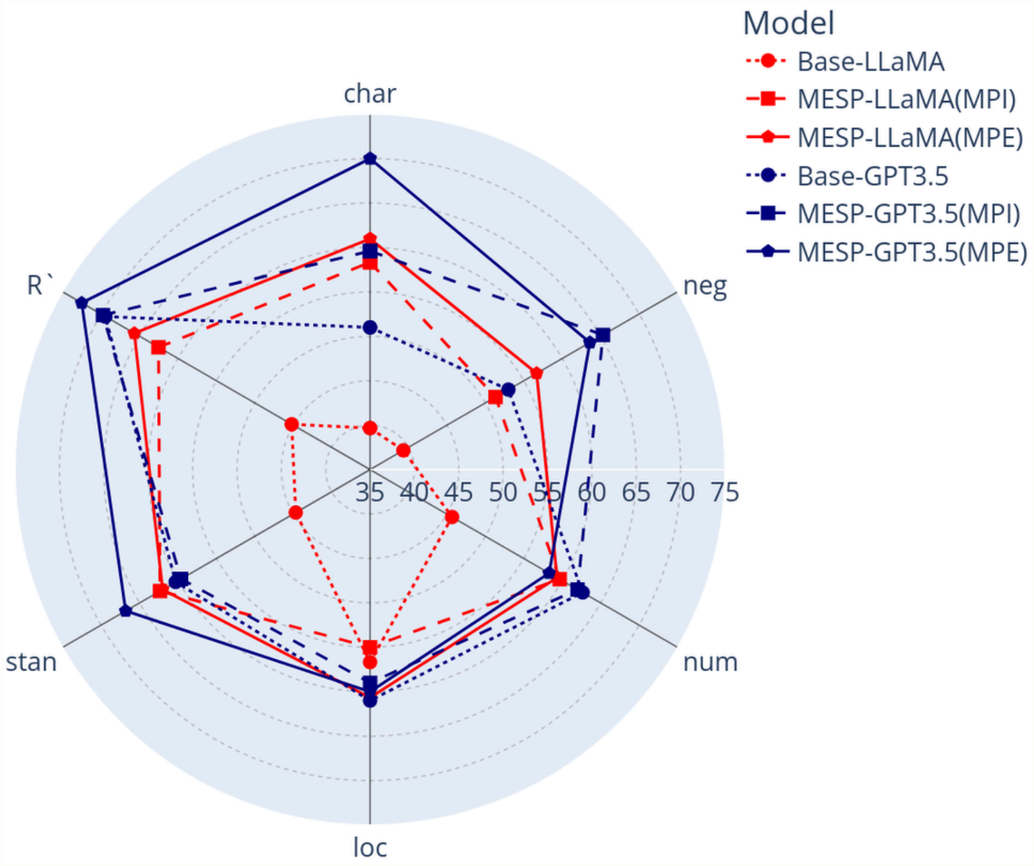
- Incorporating an input perturbation explanation within the prompt enhances the model's accuracy.

EMNLP 2024

# RESULTS AND ANALYSIS - LLMS [LARGE LANGUAGE MODELS]

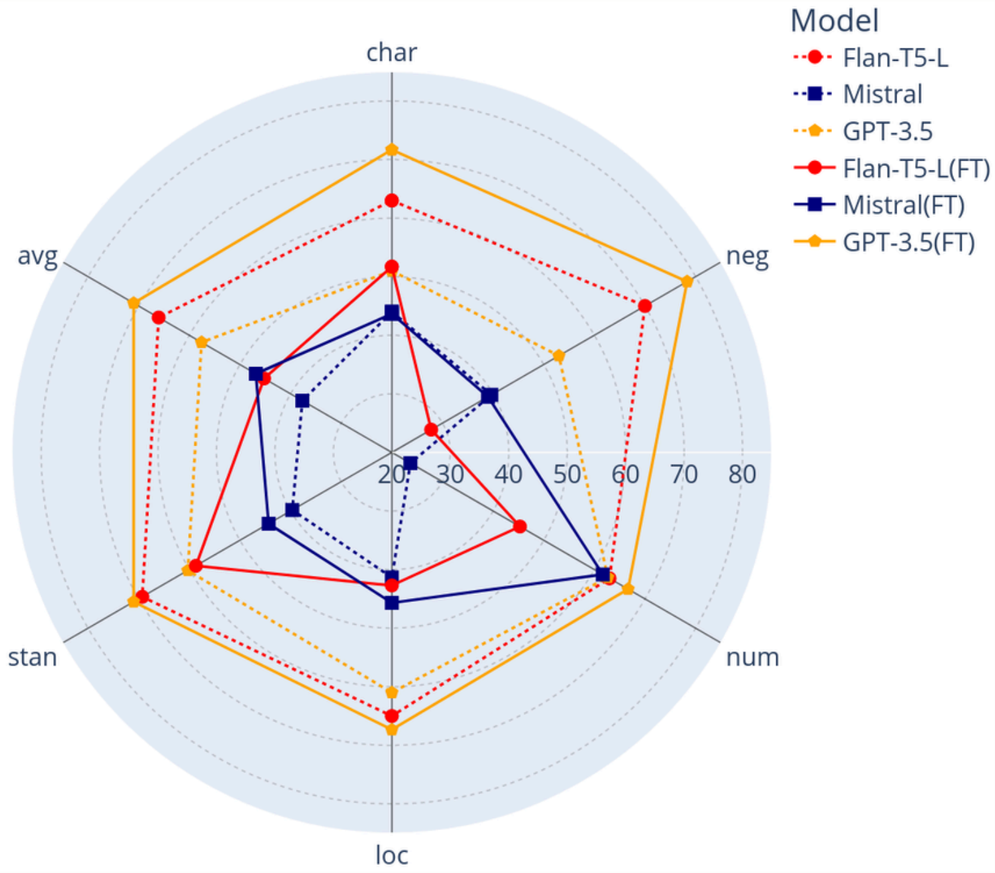
**Multiple Exemplars Single Prompts**

- LLMs, when guided with perturbation descriptions and examples, yield more stable outputs.
- Our findings show that a mixed prompting approach with several perturbation instances and brief expla- nations improves robustness.




EMNLP 2024

# RESULTS AND ANALYSIS - LLMS [LARGE LANGUAGE MODELS]




## Fine-Tuning LLMs

- For Mistral and GPT-3.5 the fine-tuning with the perturbation set using the mix training approach increases the models' performance.
- For Flan-T5-L model the fine tuning does not improve the model's performance.

EMNLP 2024

# CONCLUSION

- **Difficiulty of Concurrent Robustness** is demonstrated and it is hence shown that input perturbation poses difficulties for LMs at all scales.

- **We introduce the comprehensive Multi-Set inoculation framework** to systematically evaluate LM robustness against multiple input perturbations.

- Our results underscore the superiority of mixed fine-tuning in training robust models and the potential of such strategies to improve the model's perfromance on the concurrent robustenss problem.

EMNLP 2024

# FUTURE DIRECTIONS

- **Complex Sample Selection :** We can adopt more advanced sample selection strategies to boost model robustness during fine-tuning. This can include DataSet-Cartography[1] to figure out hard examples during training or selecting having an unbiased selection of samples for robust training [2]

- **Composite Perturbation effect:** We aim to explore the successive application of multiple perturbations on a single sample, represented as πi(πj(x)), to understand their combined impact on model performance [3]

*Note: πi represents application of a particular pertrubation (of say type i)*

Reference:
[1] Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9275–9293, Online. Association for Computa- tional Linguistics.
[2] Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. 2021. Sample selection for fair and robust training.
[3] Abhilash Shankarampeta, Vivek Gupta, and Shuo Zhang. 2022. Enhancing tabular reasoning with pattern exploiting training. In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Asso- ciation for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 706–726, Online only. Association for Computational Linguistics.

EMNLP 2024

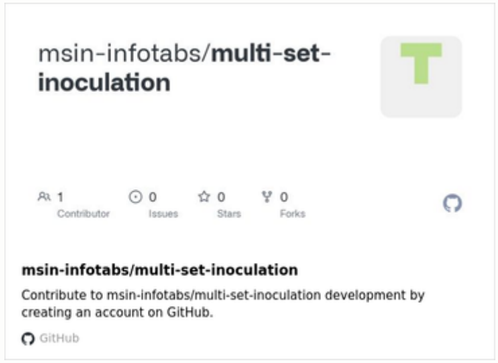
# THANK YOU!

- We would be happy to discuss and address any questions.

## MAIL

g.vatsal@iitg.ac.in
p.pandya@iitg.ac.in

## GITHUB




https://github.com/msin-infotabs/multi-set-inoculation

## PAPER

https://arxiv.org/abs/2311.08662

## WEBSITE

https://msin-infotabs.github.io

EMNLP 2024