Evaluating Concurrent Robustness of Language Models Across Diverse Challenge Sets

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

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

1. Motivation

Example Premise-Hypothesis pairs

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)

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

Figure: **An example of tabular premise and hypotheses from INFOTABS**

Original hypotheses (H_1,H_2,H_3,H_4,H_5) and perturbed hypothesis (H_1',H_2',H_3',H_4',H_5') representing character, negation, paraphrasing, numeric and location perturbations respectively. Labelled as Entailment, Contradiction or Neutral.

- **Robustness Concerns** LMs are highly sensitive to input perturbations, limiting real-world usability
- **Evaluation Gaps**: Existing analyses focus on single perturbations, but real-world inputs involve multiple, concurrent perturbations.

2. Our Contributions

- **Multi-Set Inoculation Framework**: Proposed a generalizable evaluation approach to gauge model robustness against diverse concurrent perturbations and demonstrated effectiveness on Tabular-NLI tasks with multiple perturbations.
- **Covers PLMs and LLMs**: Developed fine-tuning methods for PLMs and prompting strategies for LLMs.
- **Comprehensive Analysis**: Extensive analysis to out-of-domain (*Un-Seen*) perturbations and fine-tuning experiments for LLMs.

3. Methodology

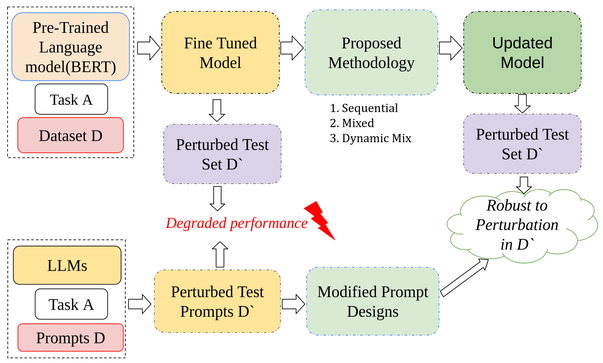

Figure: The **Multi-Set Inoculation Framework**

- **Fine-Tuning Strategies for PLMs**:
  - **Sequential Training**: Fine-tuning models on each perturbation type in succession (*sequence*).
  - **Mixed Training**: Aggregates samples from all perturbation types into a single training set.
  - **Dynamic Mixed Training**: Adjusts number of samples dynamically for fine-tuning based on model performance on each perturbation, emphasizing harder perturbations.
- **Prompting for LLMs**: We leverage Chain-of-Thought grounded prompting with exemplars to enhance concurrent robustness drawing parallels to strategies proposed for PLMs.

4. Tabular NLI-case study

Evaluation Framework

- The models are evaluated on INFOTABS dataset with various perturbations applied to hypotheses.
- 5 perturbations are analyzed: ***char:*** character swaps, ***stan:*** paraphrasing, ***neg:*** negation, ***num:*** numeric alterations, and ***loc:*** location changes.
- We compare model performance across both the original and perturbed datasets using accuracy as the primary metric (*same as micro-F1 here*).

5. Results and Discussion

(*Q1*) Do input perturbations pose a challenge for Language Models (PLMs and LLMs)?

- Yes, The model performs better on the perturbation it is fine tuned on but not on others.

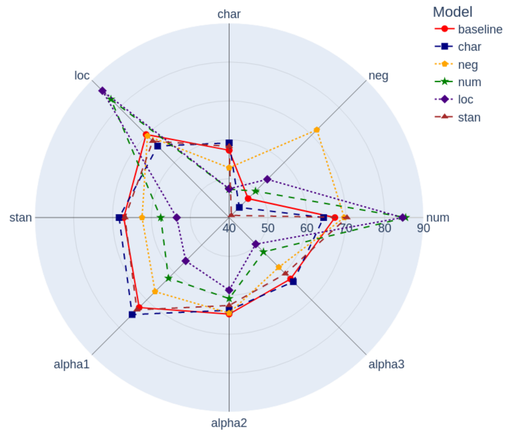

Figure: Performance of model on different perturbations when fine-tuned on a single perturbation (shown in legend)

(*Q2*) Is there a trade-off between robustness across perturbations and specialization on one type?

- Yes, specialization on one perturbation (e.g., character) often results in a performance drop on others, highlighting the importance of multi-set inoculation.

(*Q3*) How does Multi-Set Inoculation compare to single-set fine-tuning?

- **Multi-Set Inoculation** achieves better average performance across multiple perturbations compared to single-set fine-tuning. It balances between specialization and robustness across diverse input noise.

(*Q4*) Does perturbation aware prompting improve the robustness of LLMs?

- Yes, prompting LLMs to with examples of possible perturbation improves robustness significantly by making LLMs aware about the possible perturbations. Particularly, **MESP** (Multiple Exemplars Single Prompt) strategies show strong gains across all perturbation types.

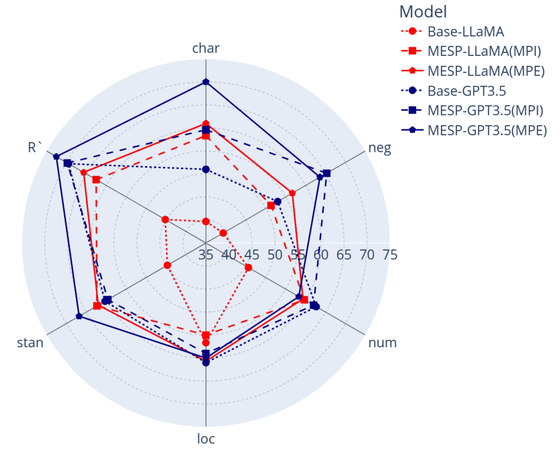

Figure: MESP Results on LLaMA-2-13b and GPT-3.5

6. Future Directions

- **Complex Sample Selection** sfor enhanced model robustness during fine-tuning
- **Multiple successive perturbations** to a single sample $[\pi_i(\pi_j(x))]$ to assess their combined effect on model performance

Figure: **Scan QR code for paper *t.ly/frRtK***

Corresponding emails
g.vatsal@iitg.ac.in, p.pandya@iitg.ac.in

Website
msin-infotabs.github.io