Prompt Perturbation and Robustness Evaluation

Shuyue Jia

Ph.D. Student

Boston University

October 3rd, 2023

Dependable Computing Laboratory,
Department of Electrical and Computer Engineering,
Boston University

Outline

- **Probing** *vs.* Prompting
- Prompt Perturbation **Category**
- Prompt Perturbation **Selected Works**
- Robustness **Problem Formulation**
- Robustness **Evaluation**

Part 1 – Probing *vs.* Prompting

- **Prompting**: use natural language to **query the LLMs** with descriptions, instructions, goals, and examples.
- The way we **access** and **interact** with a language model.

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```

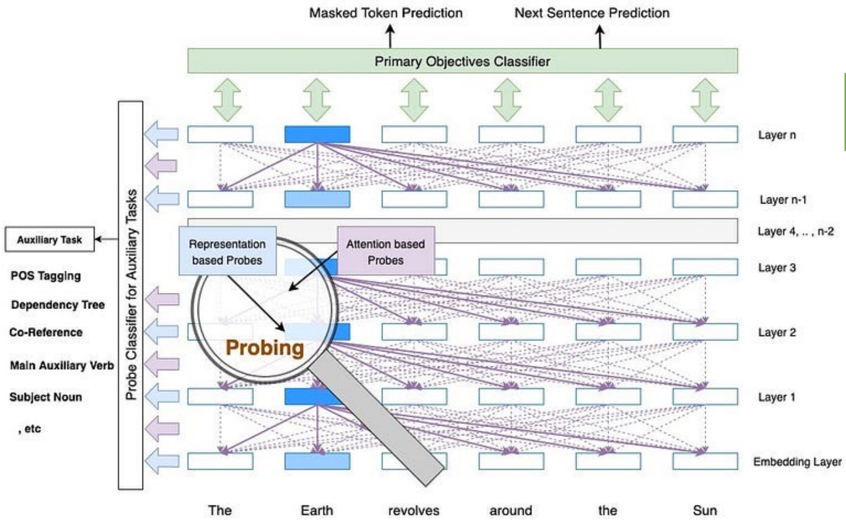
instructions

examples

goals

Image Credits: In the public domain.

Part 1 – Probing *vs.* Prompting



	Probing	Fine-tuning	Multi-task Learning
Goal	Auxiliary Task	Primary Task	Primary Tasks
Update Model Parameters	No	Yes	Yes
Access Model Internals	Yes	No	No
Complexity	Shallow	Shallow or Deep	Shallow or Deep

- **Probing**: the process of **exploring what knowledge is encoded** in the LLMs
- **Probing classifier** (diagnostic classifier) and **linear probing** (linear head)
- **Representation-based** – **Internal representation**: different layers
- **Attention-based** – **Attention Weights**

Images Credits: Medium Post – Linguistics Wisdom of NLP Models.

Part 2 – Prompt Perturbation Category

- **Prompt Perturbation**: alter or modify the original input prompt to generate semantics-preserving or varied responses.
- **Category**: different **Granularities** + **Severities**

  (1) **Character-level** – **Character Editing**

  Character swapping ("place" ⇒ "plcae"), deletion ("artist" ⇒ "arist"), insertion ("computer" ⇒ "comnputer"), substitution ("computer" ⇒ "computor"), and many more.

  (2) **Word-level** – **Word Manipulation**

  (3) **Sentence-level** – **Paraphrasing and Style Transformation**

  (4) **Adversary-level** – **Universal Adversarial Perturbation**

  Small and carefully crafted **changes**/perturbations that can be added to various input data to **cause machine learning models to make errors, *e.g.*, misclassify input text**

Credits: Li *et al.*, A Survey on Out-of-Distribution Evaluation of Neural NLP Models, In arXiv'23.

Part 2 – Prompt Perturbation Category

- **Different granularities** – **Character-level** – Character Editing
- The process of **making changes to characters** in a text.
- **Character Substituting/Replacing**, **Deleting**, **Inserting**, or **Swapping** individual characters, **Keyboard Typos** (**Typos and Misspellings**), **Optical Character Recognition** (**OCR**), and **Adding or Removing Special Symbols**.

Perturbation	Description
Character Replacement (CR)	Substitute character randomly with probability p.
Character Deletion (CD)	Delete character randomly with probability p.
Character Insertion (CI)	Insert character randomly with probability p.
Character Swap (CS)	Swap character randomly with probability p.
Keyboard Typos	Substitute character by keyboard distance with probability p.
Optical Character Recognition (OCR)	Substitute character by pre-defined OCR error with probability p.
Special Symbols Inserting or Deletion	Insert or delete Special Symbols randomly with probability p.

Credits: Qiu *et al.*, Are Multimodal Models Robust to Image and Text Perturbations?, In arXiv'23.

Part 2 – Prompt Perturbation Category

- **Different granularities** – **Character-level** – Character Editing
- The process of **making changes to characters** in a text. It involves **substituting/replacing**, **deleting**, **inserting**, or **swapping** individual characters, **keyboard typos**, optical character recognition (**OCR**), and **Adding or Removing Special Symbols**.

Perturbation	Example
Clean	**An orange metal bowl strainer filled with apples.**
Character Replacement (CR)	An orange metal towl strainer fillet with apples.
Character Deletion (CD)	An orang[X] metal bowl strainer fil[X]ed with apples.
Character Insertion (CI)	And orange metal bowl strainer filled with atpples.
Character Swap (CS)	An orange meatl bowl stariner filled with apples.
Keyboard Typos	An orange metal bowk strainer filled witj apples.
Optical Character Recognition (OCR)	An 0range metal bowl strainer filled with app1es.
Special Symbols Inserting or Deletion	An orange metal bowl? strainer filled with apples!

Credits: Qiu *et al.*, Are Multimodal Models Robust to Image and Text Perturbations?, In arXiv'23.

Part 2 – Prompt Perturbation Category

- **Different granularities** – **Word-level** – Word Manipulation
- Words are replaced with other related words, *e.g.*, **synonym replacement** (SR), **word insertion** (WR), **word swap** (WS), **word deletion** (WD), and **insert punctuation** (IP)

Perturbation	Description
Synonym Replacement (SR)	Randomly choose n words from the sentence that are not stop words. Replace each of these words with one of its synonyms chosen at random.
Word Insertion (WI)	Find a random synonym of a random word in the sentence that is not a stop word. Insert that synonym into a random position in the sentence. Do this n times.
Word Swap (WS)	Randomly choose two words in the sentence and swap their positions. Do this n times.
Word Deletion (WD)	Each word in the sentence can be randomly removed with probability p.
Insert Punctuation (IP)	Random insert punctuation in the sentence with probability p.

Credits: Qiu *et al.*, Are Multimodal Models Robust to Image and Text Perturbations?, In arXiv'23.

Part 2 – Prompt Perturbation Category

- **Different granularities** – **Word-level** – Word Manipulation
- Words are replaced with other related words, *e.g.*, **synonym replacement** (SR), **word insertion** (WR), **word swap** (WS), **word deletion** (WD), and **insert punctuation** (IP)

Perturbation	Example
Clean	**An orange metal bowl strainer filled with apples.**
Synonym Replacement (SR)	An orange alloy bowl strainer filled with apples.
Word Insertion (WI)	An old orange metal bowl strainer filled with apples.
Word Swap (WS)	An orange metal strainer bowl filled with apples.
Word Deletion (WD)	An orange metal bowl strainer [X] with apples.
Insert Punctuation (IP)	An orange metal bowl ? strainer filled with apples.

Credits: Qiu *et al.*, Are Multimodal Models Robust to Image and Text Perturbations?, In arXiv'23.

Part 2 – Prompt Perturbation Category

- **Different granularities** – **Sentence-level** – Paraphrasing
- Generates a lot of similar prompts (**paraphrasing**)
- Transfer the style of text into **formal**, **casual**, **passive**, and **active**

Perturbation	Description
Paraphrasing	Paraphrase the prompt, *e.g.*, Back Translation: Translate the source to German and translate it back to English.
Formal Style	Transfer the text style to Formal.
Casual Style	Transfer the text style to Casual.
Passive Style	Transfer the text style to Passive.
Active Style	Transfer the text style to Active.

Credits: Qiu *et al.*, Are Multimodal Models Robust to Image and Text Perturbations?, In arXiv'23.

Part 2 – Prompt Perturbation Category

- **Different granularities** – **Sentence-level** – Paraphrasing
- Generates a lot of similar prompts (**paraphrasing**)
- Transfer the style of text into **formal**, **casual**, **passive**, and **active** styles

Perturbation	Example
Clean	**An orange metal bowl strainer filled with apples.**
Paraphrasing	Apples are placed in an orange metal bowl strainer.
Formal Style	An orange metal bowl strainer contains apples.
Casual Style	An orange metal bowl is filled with apples.
Passive Style	Some apples are in an orange metal bowl strainer.
Active Style	There are apples in an orange metal bowl strainer.

Credits: Qiu *et al.*, Are Multimodal Models Robust to Image and Text Perturbations?, In arXiv'23.

Part 2 – Prompt Perturbation Category

- **Different granularities** – **Adversary-level** – Universal Adversarial Perturbation [1]
- **Universal adversarial triggers**: input-agnostic sequences of tokens that **trigger** a model to produce a specific prediction when concatenated to any input from a dataset [2].

Input (red = trigger)	Model Prediction
zoning tapping fiennes Visually imaginative, thematically instructive and thoroughly delightful, it takes us on a roller-coaster ride. . .	Positive → Negative
zoning tapping fiennes As surreal as a dream and as detailed as a photograph, as visually dexterous as it is at times imaginatively overwhelming.	Positive → Negative

Credits:
[1] Moosavi-Dezfooli *et al.*, Universal Adversarial Perturbations, In CVPR'17.
[2] Wallace *et al.*, Universal Adversarial Triggers for Attacking and Analyzing NLP, In arXiv'23.

Part 2 – Prompt Perturbation Summary

- **Category**: different **Granularities** + **Severities**
- (1) **Character-level** – **Character Editing** **(7)**
  - **Character Replacement** (CR), **Character Deletion** (CD), **Character Insertion** (CI), **Character Swap** (CS), **Keyboard Typos** (KT)**, Optical Character Recognition** (OCR), **Special Symbols Insertion or Deletion** (SS)
- (2) **Word-level** – **Word Manipulation** **(5)**
  - **Synonym Replacement** (SR), **Word Insertion** (WR), **Word Swap** (WS), **Word Deletion** (WD), **Insert Punctuation** (IP)
- (3) **Sentence-level** – **Paraphrasing and Style Transformation** **(5)**
  - **Paraphrasing** (PP)**, Formal Style** (FS)**, Casual Style** (CAS)**, Passive Style** (PS)**, Active Style** (AS)
- (4) **Adversary-level** – **Universal Adversarial Perturbation** **(1)**
  - **Universal Adversarial Triggers** (UAT)

Part 2 – Prompt Perturbation Summary

Recall@K: how many relevant items were returned *in the first K items* against how many relevant items exist in the entire dataset (TP+FN); **RSUM**: the sum of recall R@K metric

Table 23. ViLT text perturbation performance comparison of Fine-tuned (FT) image-text retrieval on Flickr30K and COCO datasets (results are averaged on five perturbation levels).

Flickr30K/COCO dataset: **1,000/5,000** images, each with **5** corresponding captions

	Method	Flickr30K (1K)									MSCOCO (5K)								
		Text Retrieval				Image Retrieval					Text Retrieval				Image Retrieval				
		R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM	R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM
Character	Keyboard	55.6	82.9	89.3	75.9	31.8	57.7	68.0	52.5	385.3	40.3	69.6	79.9	63.3	23.1	47.3	59.0	43.1	319.2
	Ocr	71.1	92.0	96.1	86.4	45.8	74.1	82.8	67.6	462.0	51.9	80.1	88.5	73.5	32.5	60.8	72.5	55.2	386.2
	CI	55.3	83.2	90.1	76.2	31.9	58.5	68.9	53.1	388.0	41.1	70.8	81.4	64.4	24.0	48.9	60.8	44.6	327.0
	CR	55.7	82.5	90.1	76.1	31.8	57.7	68.3	52.6	386.2	40.8	69.8	80.5	63.7	23.5	47.7	59.4	43.5	321.7
	CS	57.6	83.8	90.7	77.4	33.7	59.8	70.0	54.5	395.6	42.3	72.2	82.0	65.5	24.9	49.9	61.7	45.5	333.1
	CD	57.3	84.0	90.8	77.4	34.6	60.9	71.0	55.5	398.6	42.3	71.9	82.3	65.5	25.1	50.3	62.3	45.9	334.1
Word	SR	71.0	92.4	96.1	86.5	48.9	77.4	86.0	70.8	471.9	52.8	80.9	88.9	74.2	35.2	64.3	75.7	58.4	397.8
	WI	75.0	94.0	97.3	88.8	53.9	82.4	89.5	75.3	492.2	56.5	83.4	90.9	76.9	38.6	68.4	79.7	62.2	417.5
	WS	71.6	93.0	96.8	87.1	50.4	80.2	88.1	72.9	480.1	53.7	81.4	89.5	74.9	35.8	66.0	78.0	60.0	404.4
	WD	74.3	93.9	97.3	88.5	53.0	82.0	89.3	74.8	489.8	55.6	82.5	90.3	76.2	37.8	68.0	79.4	61.7	413.6
	IP	79.5	95.7	98.0	91.1	58.1	85.0	91.3	78.1	507.7	59.9	85.4	92.0	79.1	41.8	71.6	82.3	65.2	433.1
Sentence	Formal	79.5	95.7	98.6	91.3	59.2	85.6	91.5	78.8	510.1	61.1	85.8	92.2	79.7	42.6	72.2	82.6	65.8	436.5
	Casual	78.1	95.5	97.8	90.5	57.3	84.9	90.9	77.7	504.5	60.0	85.5	91.7	79.1	42.2	71.9	82.4	65.5	433.6
	Passive	74.0	94.6	97.4	88.7	53.2	80.8	88.1	74.0	488.1	57.9	84.4	91.4	77.9	40.0	69.3	80.2	63.2	423.2
	Active	78.5	95.1	98.3	90.6	58.6	85.7	92.1	78.8	508.3	60.9	85.9	92.2	79.7	42.9	72.3	82.9	66.0	437.1
	Back_trans	78.0	94.8	98.0	90.3	56.1	83.0	90.2	76.4	500.1	59.1	84.4	91.3	78.3	40.5	69.9	80.7	63.7	426.0

Credits: Qiu *et al.*, Are Multimodal Models Robust to Image and Text Perturbations?, In arXiv'23.

Part 2 – Prompt Perturbation Summary

Table 24. CLIP text perturbation performance comparison of Zero-Shot (ZS) image-text retrieval on Flickr30K and COCO datasets (results are averaged on five perturbation levels).

	Method	Flickr30K (1K)									MSCOCO (5K)								
		Text Retrieval				Image Retrieval					Text Retrieval				Image Retrieval				
		R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM	R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM
Character	Keyboard	62.4	86.9	93.1	80.8	43.5	68.8	77.0	63.1	431.8	36.8	62.1	72.8	57.2	21.0	41.2	51.6	37.9	285.5
	Ocr	73.4	93.2	96.7	87.8	52.9	77.3	84.6	71.6	478.2	37.2	62.2	72.6	57.4	21.1	41.5	51.8	38.1	286.4
	CI	66.4	89.6	94.7	83.6	47.3	72.3	80.2	66.6	450.5	37.0	62.1	72.8	57.3	21.2	41.4	51.6	38.1	286.1
	CR	63.0	88.4	93.8	81.7	44.1	68.7	77.2	63.3	435.2	36.6	62.1	72.7	57.1	21.0	41.4	51.7	38.0	285.4
	CS	65.5	89.3	94.9	83.2	45.7	70.4	78.7	65.0	444.6	36.5	62.2	72.6	57.1	21.1	41.4	51.8	38.1	285.6
	CD	66.3	90.4	95.4	84.0	47.2	71.9	80.1	66.4	451.3	36.6	62.2	73.0	57.3	21.1	41.4	51.6	38.0	285.8
Word	SR	76.0	95.1	98.0	89.7	58.0	81.7	88.2	76.0	497.1	47.0	72.8	81.8	67.2	29.2	53.0	63.6	48.6	347.5
	WI	78.3	95.7	98.3	90.8	61.6	84.9	90.9	79.1	509.6	49.9	74.9	83.5	69.4	32.1	56.5	66.9	51.8	363.8
	WS	77.2	95.1	98.0	90.1	59.7	83.6	89.8	77.7	503.3	48.9	73.6	82.3	68.3	30.6	54.7	65.3	50.2	355.5
	WD	80.9	96.8	98.5	92.1	61.4	85.4	91.1	79.3	514.1	51.7	76.4	84.6	70.9	32.3	56.5	67.1	51.9	368.6
	IP	81.8	97.1	98.8	92.6	63.8	86.1	91.6	80.5	519.4	52.4	76.6	84.5	71.2	34.1	58.2	68.4	53.6	374.2
Sentence	Formal	86.4	98.6	99.1	94.7	66.0	88.5	93.1	82.5	531.7	56.8	80.4	87.7	75.0	36.4	60.9	70.8	56.0	393.0
	Casual	84.9	97.9	99.2	94.0	66.1	88.4	92.8	82.4	529.3	57.1	79.6	87.7	74.8	35.9	60.6	70.7	55.7	391.6
	Passive	84.3	96.9	99.2	93.5	64.8	87.3	92.2	81.5	524.8	54.3	77.8	86.1	72.7	34.1	58.4	68.9	53.8	379.6
	Active	85.6	97.9	99.2	94.2	66.9	88.8	93.1	82.9	531.4	57.5	80.3	87.9	75.2	36.1	60.8	70.9	55.9	393.5
	Back_trans	83.9	97.0	98.5	93.1	65.5	87.2	92.2	81.6	524.2	55.1	78.2	85.7	73.0	34.3	58.9	69.1	54.1	381.2

Credits: Qiu *et al.*, Are Multimodal Models Robust to Image and Text Perturbations?, In arXiv'23.

Part 2 – Prompt Perturbation Summary

Table 25. CLIP text perturbation performance comparison of Fine-tuned (FT) image-text retrieval on Flickr30K and COCO datasets (results are averaged on five perturbation levels).

	Method	Flickr30K (1K)									MSCOCO (5K)								
		Text Retrieval				Image Retrieval					Text Retrieval				Image Retrieval				
		R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM	R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM
Character	Keyboard	67.0	91.2	96.2	84.8	48.3	74.0	81.6	68.0	458.4	36.8	66.1	78.1	60.3	24.3	49.4	61.3	45.0	316.1
	Ocr	76.2	95.4	98.4	90.0	58.5	83.3	89.1	77.0	500.9	36.8	66.3	77.9	60.4	24.4	49.7	61.5	45.2	316.7
	CI	71.4	93.3	96.8	87.2	53.2	78.1	84.8	72.0	477.6	36.3	66.6	78.2	60.4	24.4	49.6	61.4	45.1	316.5
	CR	68.9	91.7	96.1	85.6	48.7	74.5	81.7	68.3	461.6	36.5	66.3	78.1	60.3	24.3	49.7	61.5	45.2	316.4
	CS	70.7	92.4	96.6	86.6	51.0	76.6	83.7	70.4	471.1	36.5	66.5	78.2	60.4	24.4	49.6	61.4	45.1	316.7
	CD	70.9	93.3	97.2	87.2	52.1	77.5	84.5	71.3	475.5	36.7	66.1	77.9	60.3	24.2	49.5	61.3	45.0	315.6
Word	SR	78.0	96.4	98.5	91.0	63.4	87.2	92.0	80.9	515.4	45.3	75.0	85.1	68.5	33.8	62.7	74.3	56.9	376.2
	WI	81.0	97.0	99.0	92.3	68.3	90.4	94.7	84.4	530.4	48.4	77.3	86.8	70.8	37.3	66.8	78.1	60.7	394.6
	WS	80.8	97.0	99.0	92.2	66.1	89.3	93.9	83.1	526.0	48.0	77.1	86.7	70.6	35.9	65.3	76.9	59.4	389.9
	WD	81.0	97.4	99.1	92.5	67.9	90.7	95.0	84.5	531.1	49.1	77.7	86.8	71.2	37.1	66.7	78.0	60.6	395.3
	IP	83.0	97.9	99.2	93.4	69.9	91.2	95.1	85.4	536.4	51.5	79.5	88.1	73.0	39.1	68.7	79.6	62.5	406.6
Sentence	Formal	85.2	98.4	99.5	94.4	73.3	92.9	96.4	87.6	545.8	53.5	81.0	88.9	74.5	41.7	70.8	81.3	64.6	417.3
	Casual	83.9	97.6	99.4	93.6	72.5	92.3	96.4	87.1	542.1	52.5	80.6	89.0	74.0	41.4	70.4	81.2	64.4	415.2
	Passive	82.9	97.7	99.1	93.2	71.3	91.3	95.6	86.1	537.9	51.9	80.0	88.3	73.4	39.6	68.9	80.0	62.8	408.7
	Active	85.0	97.6	99.4	94.0	73.5	92.9	96.6	87.7	545.1	54.1	81.4	89.0	74.8	42.2	71.1	81.7	65.0	419.4
	Back_trans	83.8	97.7	99.0	93.5	70.4	91.2	95.2	85.6	537.3	51.4	79.1	88.2	72.9	39.6	68.5	79.5	62.5	406.2

Credits: Qiu *et al.*, Are Multimodal Models Robust to Image and Text Perturbations?, In arXiv'23.

Part 2 – Prompt Perturbation Summary

Table 26. BLIP text perturbation performance comparison of Fine-tuned (FT) image-text retrieval on Flickr30K and COCO datasets (results are averaged on five perturbation levels).

	Method	Flickr30K (1K)									MSCOCO (5K)								
		Text Retrieval				Image Retrieval					Text Retrieval				Image Retrieval				
		R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM	R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM
Character	Keyboard	84.5	97.3	98.9	93.6	63.8	84.1	89.4	79.1	518.0	64.1	86.4	91.9	80.8	42.7	67.5	76.6	62.2	429.1
	Ocr	93.6	99.5	99.8	97.6	77.5	93.1	96.0	88.9	559.5	74.3	92.2	96.0	87.5	53.6	77.7	85.3	72.2	479.1
	CI	86.6	98.0	99.3	94.7	66.3	86.1	90.9	81.1	527.3	66.7	88.1	93.4	82.7	45.0	70.2	79.0	64.7	442.4
	CR	84.6	97.5	99.0	93.7	63.9	83.8	89.2	79.0	518.0	64.5	86.7	92.1	81.1	42.9	67.7	76.9	62.5	430.8
	CS	87.4	97.9	99.3	94.9	65.9	85.4	90.5	80.6	526.4	67.0	88.1	93.2	82.8	44.6	69.7	78.6	64.3	441.3
	CD	86.8	97.7	99.2	94.6	65.9	85.7	90.4	80.7	525.7	67.0	88.1	93.3	82.8	44.8	69.7	78.6	64.4	441.4
Word	SR	93.8	99.6	99.9	97.8	80.6	94.7	97.0	90.7	565.6	74.2	92.4	96.1	87.6	55.5	79.5	86.7	73.9	484.3
	WI	96.0	99.8	99.9	98.6	85.0	96.9	98.5	93.4	576.1	78.1	94.0	97.1	89.7	60.1	83.2	89.6	77.6	502.1
	WS	94.8	99.6	100.0	98.1	83.6	96.5	98.4	92.8	572.9	75.9	93.2	96.6	88.6	58.1	82.0	88.9	76.3	494.6
	WD	95.1	99.8	100.0	98.3	83.8	96.7	98.5	93.0	573.8	77.3	93.9	97.0	89.4	59.2	82.7	89.5	77.1	499.7
	IP	97.3	99.9	100.0	99.0	87.2	97.5	98.9	94.5	580.7	81.8	95.4	97.8	91.7	63.9	85.6	91.3	80.3	515.8
Sentence	Formal	96.5	99.9	100.0	98.8	86.7	97.1	98.8	94.2	579.0	81.7	95.2	97.6	91.5	63.5	85.3	91.2	80.0	514.4
	Casual	96.8	100.0	100.0	98.9	86.0	97.1	98.7	93.9	578.6	81.3	95.0	97.7	91.3	63.4	85.1	91.1	79.8	513.6
	Passive	96.8	99.8	99.9	98.8	83.3	96.5	98.2	92.7	574.5	80.5	94.7	97.3	90.8	61.7	83.8	90.2	78.6	508.1
	Active	97.1	99.9	100.0	99.0	86.6	97.2	98.7	94.2	579.6	81.6	95.2	97.7	91.5	64.0	85.5	91.3	80.3	515.4
	Back_trans	96.0	99.9	100.0	98.6	84.5	96.1	98.2	92.9	574.7	79.9	94.2	97.0	90.4	61.0	82.9	89.3	77.8	504.3

Credits: Qiu *et al.*, Are Multimodal Models Robust to Image and Text Perturbations?, In arXiv'23.

Part 2 – Prompt Perturbation Summary

Table 27. ALBEF text perturbation performance comparison of Fine-tuned (FT) image-text retrieval on Flickr30K and COCO datasets (results are averaged on five perturbation levels).

	Method	Flickr30K (1K)									MSCOCO (5K)								
		Text Retrieval				Image Retrieval					Text Retrieval				Image Retrieval				
		R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM	R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM
Character	Keyboard	82.1	96.0	98.5	92.2	59.7	82.1	87.7	76.5	506.2	57.9	82.6	89.6	76.7	38.0	63.4	73.0	58.1	404.5
	Ocr	91.3	99.2	99.6	96.7	74.6	92.1	95.1	87.3	552.0	69.3	89.9	94.8	84.7	49.5	74.9	83.3	69.2	461.7
	CI	84.4	97.2	98.6	93.4	62.5	84.2	89.2	78.6	516.2	60.8	84.7	91.0	78.8	40.6	66.2	75.6	60.8	418.9
	CR	82.1	95.9	98.4	92.1	59.9	81.6	87.2	76.2	505.0	58.3	82.9	89.9	77.0	38.3	63.6	73.1	58.3	406.1
	CS	82.9	96.8	98.8	92.8	61.6	83.2	88.4	77.7	511.7	59.9	84.1	90.8	78.3	39.8	65.3	74.8	60.0	414.7
	CD	83.6	96.7	98.5	92.9	61.9	83.6	88.7	78.1	513.0	60.0	84.1	90.8	78.3	39.9	65.7	75.1	60.2	415.5
Word	SR	92.9	99.2	99.8	97.3	78.7	94.5	96.8	90.0	561.9	70.1	90.6	95.1	85.3	52.4	77.7	85.5	71.9	471.4
	WI	94.3	99.6	99.9	97.9	82.9	96.6	98.3	92.6	571.6	73.2	92.4	96.3	87.3	56.8	81.6	88.7	75.7	488.9
	WS	93.3	99.4	99.9	97.6	81.5	96.3	98.1	92.0	568.6	72.0	91.8	96.1	86.6	55.1	80.6	88.2	74.6	483.7
	WD	93.4	99.5	99.9	97.6	82.2	96.5	98.3	92.4	570.0	72.9	92.1	96.1	87.0	55.7	81.1	88.5	75.1	486.3
	IP	95.9	99.8	100.0	98.6	85.5	97.5	98.9	94.0	577.7	77.6	94.3	97.2	89.7	60.7	84.3	90.5	78.5	504.5
Sentence	Formal	95.4	99.7	99.9	98.3	85.2	97.3	98.7	93.7	576.2	77.6	94.1	97.0	89.6	60.2	83.9	90.3	78.1	503.1
	Casual	95.1	99.7	100.0	98.3	84.6	97.1	98.5	93.4	575.0	77.1	94.1	97.4	89.5	59.7	83.6	90.1	77.8	502.0
	Passive	94.6	99.4	100.0	98.0	81.5	96.1	98.0	91.8	569.5	76.1	93.4	96.7	88.7	58.4	82.6	89.2	76.7	496.4
	Active	95.6	99.8	100.0	98.5	85.0	97.3	98.7	93.7	576.4	77.5	94.2	97.1	89.6	60.4	84.2	90.3	78.3	503.7
	Back_trans	95.9	99.7	99.9	98.5	83.0	96.1	98.0	92.3	572.5	75.2	93.0	96.4	88.2	57.4	81.0	88.3	75.6	491.3

Credits: Qiu *et al.*, Are Multimodal Models Robust to Image and Text Perturbations?, In arXiv'23.

Part 2 – Prompt Perturbation Summary

Table 28. TCL text perturbation performance comparison of Zero-Shot (ZS) image-text retrieval on Flickr30K and COCO datasets (results are averaged on five perturbation levels).

	Method	Flickr30K (1K)									MSCOCO (5K)								
		Text Retrieval				Image Retrieval					Text Retrieval				Image Retrieval				
		R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM	R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM
Character	Keyboard	63.8	87.2	92.7	81.2	44.1	68.8	76.7	63.2	433.3	49.6	76.1	84.9	70.2	32.3	57.2	67.8	52.4	368.0
	Ocr	78.2	94.8	97.9	90.3	58.8	82.1	88.1	76.3	499.9	61.4	85.1	91.6	79.4	42.6	69.0	78.7	63.4	428.4
	CI	67.3	88.0	93.4	82.9	45.9	70.5	78.3	64.9	443.3	51.9	78.5	86.7	72.4	34.1	59.8	70.3	54.7	381.3
	CR	63.1	85.9	91.4	80.1	43.8	68.1	76.1	62.7	428.4	49.7	76.1	85.1	70.3	32.2	57.4	67.9	52.5	368.4
	CS	66.5	88.6	93.8	83.0	46.3	70.8	78.5	65.2	444.4	52.6	78.5	87.0	72.7	34.0	59.7	70.1	54.6	382.0
	CD	66.7	89.4	94.2	83.4	47.2	71.9	79.4	66.2	448.9	52.6	78.8	86.9	72.8	34.3	60.2	70.6	55.0	383.4
Word	SR	78.3	95.3	97.9	90.5	63.2	86.0	91.1	80.1	511.9	62.1	85.7	91.9	79.9	45.8	72.3	81.5	66.5	439.3
	WI	80.0	96.3	98.5	91.6	67.0	88.6	93.4	83.0	523.8	63.3	86.8	93.0	81.0	49.5	76.1	84.7	70.1	453.4
	WS	80.4	95.9	98.4	91.6	64.8	87.2	92.4	81.5	519.1	63.2	86.5	92.7	80.8	46.5	73.8	83.0	67.8	445.7
	WD	83.6	97.1	98.8	93.1	67.0	89.0	93.4	83.1	528.8	65.3	87.2	93.1	81.9	47.6	74.4	83.3	68.4	450.9
	IP	89.4	98.6	99.6	95.9	73.4	92.2	95.5	87.0	548.6	71.4	90.8	95.4	85.9	53.5	79.0	87.1	73.2	477.2
Sentence	Formal	88.0	98.0	99.8	95.3	72.0	91.6	95.1	86.2	544.4	70.8	90.6	95.2	85.5	52.9	78.4	86.5	72.6	474.4
	Casual	87.2	98.3	99.5	95.0	71.4	91.2	94.8	85.8	542.4	69.9	90.2	94.9	85.0	52.3	78.1	86.4	72.3	471.8
	Passive	84.5	97.1	99.4	93.7	67.6	88.6	92.9	83.0	530.1	68.6	89.1	94.4	84.0	50.5	76.9	85.2	70.9	464.7
	Active	89.3	98.3	99.9	95.8	72.9	91.5	95.1	86.5	547.1	70.9	90.6	95.3	85.6	53.1	78.9	86.9	73.0	475.7
	Back_trans	86.0	97.6	99.4	94.3	69.4	89.8	93.6	84.3	535.8	68.5	89.2	94.2	83.9	50.3	75.9	84.1	70.1	462.0

Credits: Qiu *et al.*, Are Multimodal Models Robust to Image and Text Perturbations?, In arXiv'23.

Part 2 – Prompt Perturbation Summary

Table 29. TCL text perturbation performance comparison of Fine-tuned (FT) image-text retrieval on Flickr30K and COCO datasets (results are averaged on five perturbation levels).

	Method	Flickr30K (1K)									MSCOCO (5K)								
		Text Retrieval				Image Retrieval					Text Retrieval				Image Retrieval				
		R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM	R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	RSUM
Character	Keyboard	79.7	95.2	97.9	90.9	57.0	79.1	85.4	73.8	494.3	55.8	81.3	88.8	75.3	36.9	62.5	72.4	57.3	397.8
	Ocr	90.0	99.1	99.7	96.3	71.7	90.4	94.0	85.4	545.0	67.6	88.9	94.0	83.5	48.0	73.9	82.6	68.2	455.1
	CI	82.2	96.2	98.3	92.2	59.6	81.4	87.2	76.1	504.9	58.5	83.5	90.4	77.5	39.3	65.3	75.0	59.8	412.0
	CR	79.3	94.8	97.8	90.7	56.7	79.1	85.0	73.6	492.8	55.6	81.5	89.0	75.4	37.2	62.7	72.5	57.5	398.5
	CS	80.7	96.0	98.2	91.6	59.0	81.2	86.8	75.7	501.9	57.6	82.9	90.2	76.9	38.7	64.8	74.6	59.4	408.8
	CD	81.4	95.7	98.3	91.8	59.1	81.2	86.7	75.7	502.4	58.1	83.0	90.0	77.0	39.2	65.3	75.0	59.8	410.5
Word	SR	91.0	99.1	99.7	96.6	76.1	93.0	95.8	88.3	554.7	67.8	89.1	94.2	83.7	51.0	76.8	84.8	70.8	463.7
	WI	93.4	99.4	99.8	97.5	80.5	95.5	97.7	91.2	566.4	70.8	91.0	95.6	85.8	55.3	80.6	88.0	74.6	481.3
	WS	91.0	99.1	99.6	96.6	78.2	94.7	97.4	90.1	560.0	69.2	90.3	94.9	84.8	52.3	78.5	86.6	72.5	471.8
	WD	92.6	99.4	99.8	97.3	79.5	95.3	97.6	90.8	564.2	70.8	90.7	95.5	85.7	53.7	79.7	87.3	73.6	477.7
	IP	94.9	99.5	99.8	98.1	84.0	96.7	98.5	93.1	573.4	75.6	92.8	96.7	88.3	59.0	83.2	89.9	77.3	497.1
Sentence	Formal	94.4	99.4	99.8	97.9	83.2	96.5	98.3	92.6	571.5	75.3	92.4	96.7	88.1	58.2	82.7	89.5	76.8	494.6
	Casual	94.0	99.5	99.9	97.8	82.1	96.0	98.0	92.1	569.6	74.6	92.1	96.5	87.8	57.9	82.5	89.4	76.6	493.0
	Passive	92.7	99.1	99.8	97.2	79.5	94.5	97.1	90.4	562.8	73.5	91.9	96.1	87.2	56.3	81.3	88.3	75.3	487.3
	Active	94.8	99.5	99.8	98.0	83.5	96.4	98.2	92.7	572.1	75.4	92.7	96.6	88.2	58.7	83.0	89.7	77.1	496.0
	Back_trans	93.9	99.5	99.9	97.8	80.6	95.3	97.3	91.1	566.5	72.7	91.6	96.0	86.8	55.5	80.3	87.3	74.4	483.5

Credits: Qiu *et al.*, Are Multimodal Models Robust to Image and Text Perturbations?, In arXiv'23.

Part 3 – Prompt Perturbation Selected Works

Related Work 1: Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How Can We Know What Language Models Know?, *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.

How Can We Know What Language Models Know?

Zhengbao Jiang[1*] **Frank F. Xu**[1*] **Jun Araki**[2] **Graham Neubig**[1]
Language Technologies Institute, Carnegie Mellon University[1]
Bosch Research North America[2]
{zhengbaj,fangzhex,gneubig}@cs.cmu.edu jun.araki@us.bosch.com

Abstract

Recent work has presented intriguing results examining the knowledge contained in language models (LM) by having the LM fill in the blanks of prompts such as "*Obama is a _ by profession*". These prompts are usually manually created, and quite possibly sub-optimal; another prompt such as "*Obama worked as a _*" may result in more accurately predicting the correct profession. Because of this, given an inappropriate prompt, we might fail to retrieve facts that the LM *does* know, and thus any given prompt only provides a lower bound estimate of the knowledge contained in an LM. In this paper, we attempt to more accurately estimate the knowledge contained in LMs by automati-

Prompts

manual	DirectX *is developed by* y_{man}
mined	y_{mine} *released the* DirectX
paraphrased	DirectX *is created by* y_{para}

Top 5 predictions and log probabilities

	y_{man}		y_{mine}		y_{para}	
1	Intel	-1.06	Microsoft	-1.77	Microsoft	-2.23
2	Microsoft	-2.21	They	-2.43	Intel	-2.30
3	IBM	-2.76	It	-2.80	default	-2.96
4	Google	-3.40	Sega	-3.01	Apple	-3.44
5	Nokia	-3.58	Sony	-3.19	Google	-3.45

Figure 1: Top-5 predictions and their log probabilities using different prompts (manual, mined, and paraphrased) to query BERT. Correct answer is underlined.

where the hidden vectors learned through a language modeling objective are then used in down-stream language understanding systems (Dai and

Challenges & Main ideas

1. **Manually** created prompts **sub-optimal** → **Automatically generate** high-quality and diverse prompts
2. **GPT** → **Unstable/unnatural English** → **BERT**
3. **Prompt Generation** → **Prompt Selection**
4. **Ensemble methods** to combine answers from different prompts

Part 3 – Prompt Perturbation Selected Works

Objectives

- ***Prompt Generation***
  - Mining-based Generation
  - Paraphrasing-based Generation
- ***Prompt Selection***
  - Top-1 Prompt Selection
- ***Prompt Ensembling***
  - Rank-based Ensemble
  - Optimized Ensemble

	Prompts
manual	`DirectX` *is developed by* y_{man}
mined	y_{mine} *released the* `DirectX`
paraphrased	`DirectX` *is created by* y_{para}

Top 5 predictions and log probabilities

	y_{man}		y_{mine}		y_{para}	
1	`Intel`	-1.06	<u>`Microsoft`</u>	-1.77	<u>`Microsoft`</u>	-2.23
2	<u>`Microsoft`</u>	-2.21	`They`	-2.43	`Intel`	-2.30
3	`IBM`	-2.76	`It`	-2.80	`default`	-2.96
4	`Google`	-3.40	`Sega`	-3.01	`Apple`	-3.44
5	`Nokia`	-3.58	`Sony`	-3.19	`Google`	-3.45

Figure 1: Top-5 predictions and their log probabilities using different prompts (manual, mined, and paraphrased) to query BERT. Correct answer is underlined.

Credits: Jiang *et al.*, How Can We Know What Language Models Know?, In TACL'20.

Part 3 – Prompt Perturbation Selected Works

Prompt Generation – Mining-based Generation (**diverse**)

- **Relation Triples**: Subject-Relation-Object $< x, r, y >$
- **Observation**: Words in the vicinity of the subject x and object y in a large corpus often describe the relation r
- **Method 1**: **Middle-word Prompts** → r is used as a template

Barack Obama was born in Hawaii.

Prompts	Top1	Top3	Top5	Opti.	Oracle
Mid	30.7	32.7	31.2	36.9	45.1
Mid+Dep	31.4	34.2	34.7	38.9	50.7

Table 7: Ablation study of middle-word and dependency-based prompts on BERT-base.

- **Method 2**: **Dependency Parser-based Prompts**

  Syntactic analysis of the sentence → shortest dependency path

The capital of France is Paris.

Credits: Jiang *et al.*, How Can We Know What Language Models Know?, In TACL'20.

Part 3 – Prompt Perturbation Selected Works

Prompt Generation – Paraphrasing-based Generation

- **Back Translation**
- First, translate the initial prompt into B candidates in another language, each of which is then back-translated into B candidates in the original language → B^2 prompts
- **Round-trip probability** $P_{\text{forward}}(\bar{t}|\hat{t}) \times P_{\text{backward}}(t|\bar{t})$

  $\hat{t}$: the initial prompt

  $\bar{t}$: the translated prompt in the other language

  t: the final prompt

Credits: Jiang *et al.*, How Can We Know What Language Models Know?, In TACL'20.

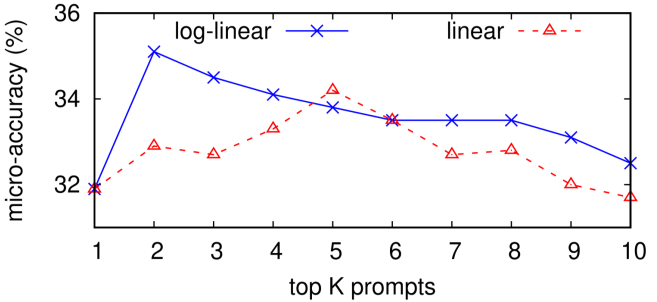
Part 3 – Prompt Perturbation Selected Works

- ***Prompt Selection***

$$A(t_{r,i}) = \frac{\sum_{<x,y>\in\mathcal{R}} \delta(y = \operatorname{argmax}_{y'} P_{\mathrm{LM}}(y'|x, t_{r,i}))}{|\mathcal{R}|},$$

$\delta(\cdot)$: Kronecker's delta function

$\mathcal{R}$: a set of subject-object pairs with relation



- ***Rank-based Ensemble***

$$s(y|x,r) = \sum_{i=1}^{K} \frac{1}{K} \log P_{\mathrm{LM}}(y|x, t_{r,i}),\ P(y|x, t_{r,i}) = \operatorname{softmax}(s(\cdot\,|x,r))_y,$$

where $t_{r,i}$ is the prompt ranked at the i-th position, and K is number

Credits: Jiang *et al.*, How Can We Know What Language Models Know?, In TACL'20.

Part 3 – Prompt Perturbation Selected Works

Data

- LAMA benchmark (**LA**nguage **M**odel **A**nalysis) [1] – T-REx subset (T-REx knowledge source) [2]: **41 relations, each with 1,000 subject-object pairs** from Wikipedia. (LAMA: probe to test the factual and commonsense knowledge: either subject-relation-object triples or question-answer pairs)
- LAMA-UHN – T-REx subset [3]: filter out those easy-to-guess facts from LAMA
- Google-RE subset (relation-extraction-corpus): **3 relations** ("place of birth", "date of birth", and "place of death"), **with ≈ 60K facts** manually extracted from Wikipedia

Models

- BERT-base and BERT-large models [4]

Model	Man	Mine	Mine +Man	Mine +Para	Man +Para
BERT	31.1	38.9	39.6	36.2	37.3
ERNIE [5]	32.1	42.3	43.8	40.1	41.1
KnowBert	26.2	34.1	34.6	31.9	32.1

Table 8: Micro-averaged accuracy (%) of various LMs

Credits:
[1] Petroni *et al.*, Language Models as Knowledge Bases?, In EMNLP'19.
[2] ElSahar *et al.*, T-REx: A Large Scale Alignment of Natural Language with Knowledge Base Triples, In LREC'18.
[3] Porner *et al.*, BERT is Not a Knowledge Base (Yet): Factual Knowledge *vs.* Name-based Reasoning in Unsupervised QA, In arXiv'20.
[4] Devlin *et al.*, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, In NAACL'19.
[5] Zhang *et al.*, ERNIE: Enhanced Language Representation with Informative Entities, In ACL'19.

Part 3 – Prompt Perturbation Selected Works

Evaluation Metrics

- **Micro-averaged Accuracy**

$$\frac{1}{|\mathcal{R}|} \sum_{<x,y> \in \mathcal{R}} \delta(\hat{y} = y) .$$

$\hat{y}$ is the prediction, and y is the ground truth.

Since object distributions of some relations are extremely skewed,

- **Macro-averaged Accuracy**

$$\frac{1}{|\mathrm{uni\_obj}(\mathcal{R})|} \sum_{y' \in \mathrm{uni\_obj}(\mathcal{R})} \frac{\sum_{<x,y> \in \mathcal{R}, y=y'} \delta(\hat{y} = y)}{|\{y| < x, y > \in \mathcal{R}, y = y'\}|},$$

where $\mathrm{uni\_obj}(\mathcal{R})$ denotes a set of unique objects from relation r.

Credits: Jiang *et al.*, How Can We Know What Language Models Know?, In TACL'20.

Part 3 – Prompt Perturbation Selected Works

Results

1. Man: lower bound **2.** Man: complicated syntactically **3.** Top-K

Prompts	Top1	Top3	Top5	Opti.	Oracle
	*BERT-base (**Man**=31.1)*				
Mine	31.4	34.2	34.7	38.9	50.7
Mine+Man	31.6	35.9	35.1	**39.6**	52.6
Mine+Para	32.7	34.0	34.5	36.2	48.1
Man+Para	*34.1*	35.8	36.6	37.3	47.9
	*BERT-large (**Man**=32.3)*				
Mine	37.0	37.0	36.4	43.7	54.4
Mine+Man	*39.4*	40.6	38.4	**43.9**	56.1
Mine+Para	37.8	38.6	38.6	40.1	51.8
Man+Para	35.9	37.3	38.0	38.8	50.0

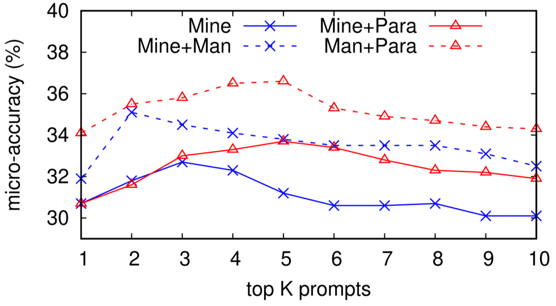
Table 2: Micro-averaged accuracy of different methods (%). **Majority** gives us 22.0%. Italic indicates best single-prompt accuracy, and bold indicates the best non-oracle accuracy overall.

Prompts	Top1	Top3	Top5	Opti.	Oracle
	*BERT-base (**Man**=22.8)*				
Mine	20.7	22.7	23.9	25.7	36.2
Mine+Man	21.3	23.8	24.8	**26.6**	38.0
Mine+Para	21.2	22.4	23.0	23.6	34.1
Man+Para	*22.8*	23.8	24.6	25.0	34.9
	*BERT-large (**Man**=25.7)*				
Mine	26.4	26.3	25.9	30.1	40.7
Mine+Man	*28.1*	28.3	27.3	**30.7**	42.2
Mine+Para	26.2	27.1	27.0	27.1	38.3
Man+Para	25.9	27.8	28.3	28.0	39.3

Table 3: Macro-averaged accuracy of different methods (%). **Majority** gives us 2.2%. Italic indicates best single-prompt accuracy, and bold indicates the best non-oracle accuracy overall.

upper bound

(somehow) lower bound



Credits: [1] Jiang *et al.*, How Can We Know What Language Models Know?, In TACL'20.
[2] **Man** (baseline) → Petroni *et al.*, Language Models as Knowledge Bases?, In EMNLP'19.

Part 3 – Prompt Perturbation Selected Works

Results

1. Man → Mine **2.** Opti+Mine **3.** Prompt Modification

ID	Relations	Manual Prompts	Mined Prompts	Acc. Gain
P140	religion	x is affiliated with the y religion	x who converted to y	+60.0
P159	headquarters location	The headquarter of x is in y	x is based in y	+4.9
P20	place of death	x died in y	x died at his home in y	+4.6
P264	record label	x is represented by music label y	x recorded for y	+17.2
P279	subclass of	x is a subclass of y	x is a type of y	+22.7
P39	position held	x has the position of y	x is elected y	+7.9

Table 4: Micro-averaged accuracy gain (%) of the mined prompts over the manual prompts.

ID	Relations	Prompts and Weights	Acc. Gain
P127	owned by	x is owned by y $_{.485}$ x was acquired by y $_{.151}$ x division of y $_{.151}$	+7.0
P140	religion	x who converted to y $_{.615}$ y tirthankara x $_{.190}$ y dedicated to x $_{.110}$	+12.2
P176	manufacturer	y introduced the x $_{.594}$ y announced the x $_{.286}$ x attributed to the y $_{.111}$	+7.0

Table 5: Weights of top-3 mined prompts, and the micro-averaged accuracy gain (%) over using the top-1 prompt.

ID	Modifications	Acc. Gain
P413	x plays in→at y position	+23.2
P495	x was created→made in y	+10.8
P495	x was→is created in y	+10.0
P361	x is a part of y	+2.7
P413	x plays ~~in~~ y position	+2.2

Table 6: Small modifications (update, insert, and delete) in paraphrase lead to large accuracy gain (%).

Credits: [1] Jiang *et al.*, How Can We Know What Language Models Know?, In TACL'20.
[2] **Man** (baseline) → Petroni *et al.*, Language Models as Knowledge Bases?, In EMNLP'19.

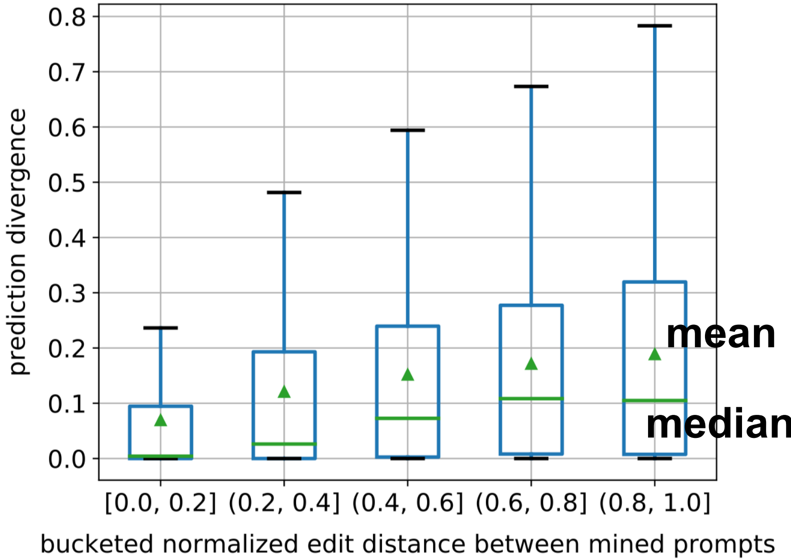
Part 3 – Prompt Perturbation Selected Works

ID	Relations	Manual Prompts	Mined Prompts	Acc. Gain
P140	religion	x is affiliated with the y religion	x who converted to y	+60.0
P159	headquarters location	The headquarter of x is in y	x is based in y	+4.9
P20	place of death	x died in y	x died at his home in y	+4.6
P264	record label	x is represented by music label y	x recorded for y	+17.2
P279	subclass of	x is a subclass of y	x is a type of y	+22.7
P39	position held	x has the position of y	x is elected y	+7.9

Table 4: Micro-averaged accuracy gain (%) of the mined prompts over the manual prompts.

ID	Relations	Prompts and Weights	Acc. Gain
P127	owned by	x is owned by y .485 x was acquired by y .151 x division of y .151	+7.0
P140	religion	x who converted to y .615 y tirthankara x .190 y dedicated to x .110	+12.2
P176	manufacturer	y introduced the x .594 y announced the x .286 x attributed to the y .111	+7.0

Table 5: Weights of top-3 mined prompts, and the micro-averaged accuracy gain (%) over using the top-1 prompt.



$$\mathrm{Div}(t_{r,i}, t_{r,j}) = \frac{\sum_{<x,y>\in\mathcal{R}} \delta(C(x,y,t_{r,i}) \neq C(x,y,t_{r,j}))}{|\mathcal{R}|}.$$

Credits:

[1] Jiang *et al.*, How Can We Know What Language Models Know?, In TACL'20.

[2] **Man** (baseline) → Petroni *et al.*, Language Models as Knowledge Bases?, In EMNLP'19.

Part 3 – Prompt Perturbation Selected Works

Limitations

- ***Scenarios***: factual knowledge extraction in the form of **relation triples**
- ***Scenarios***: limited by **relation types**
- ~~***Manual Effort***: **Manually select** a prompt from the mined set~~
- ***Prediction***: **single-token object**
- ***Generation***: Current mining-based generation is limited to **Wikipedia**
- ***Technical details*** are not revealed and open-sourced, unfortunately.

Dataset for Mining

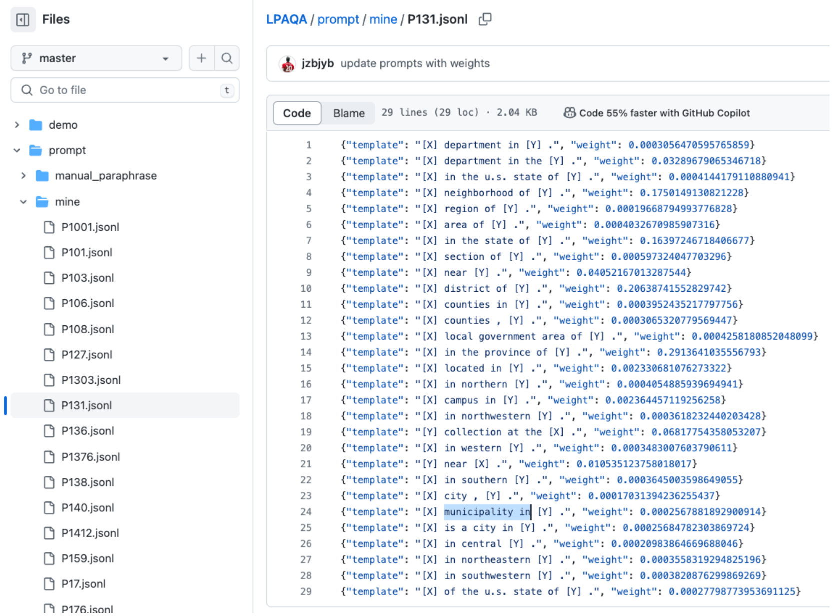
- **Wiki-ZSL (Wiki Zero-Shot Learning) dataset**: **113 relations** and **94,383 instances**

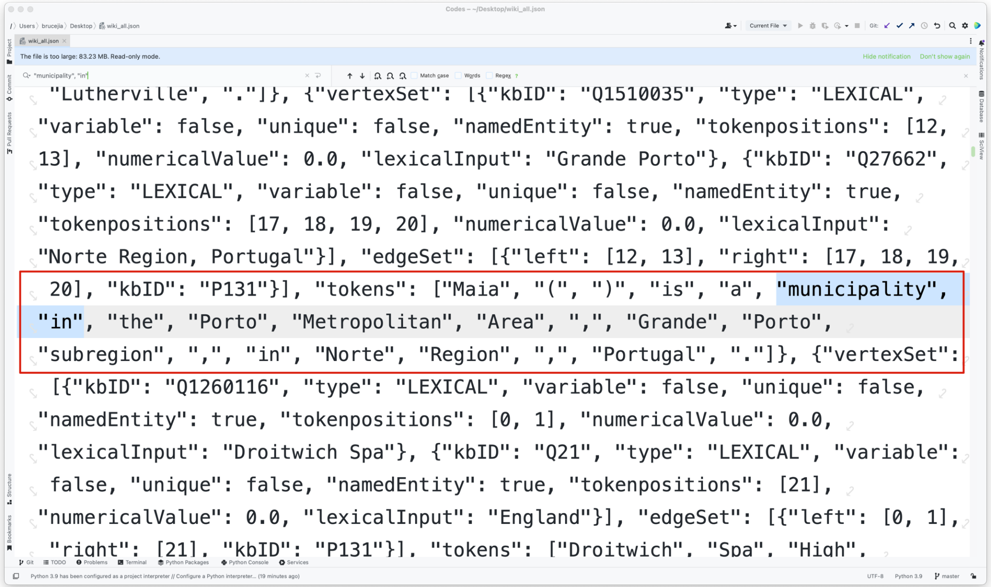
Credits: Jiang *et al.*, How Can We Know What Language Models Know?, In TACL'20.

Part 3 – Prompt Perturbation Selected Works

Dataset for Mining

- Wiki-ZSL (Wiki Zero-Shot Learning) dataset: **113 relations** and **94,383 instances**





Credits: Sorokin *et al.*, Context-Aware Representations for Knowledge Base Relation Extraction, In EMNLP'17.

Part 3 – Prompt Perturbation Selected Works

Related Work 2: Mohna Chakraborty, Adithya Kulkarni, and Qi Li.
Zero-shot Approach to Overcome Perturbation Sensitivity of Prompts,
Association for Computational Linguistics, 1:5698–5711, 2023.

Zero-shot Approach to Overcome Perturbation Sensitivity of Prompts

Mohna Chakraborty*, **Adithya Kulkarni***, and **Qi Li**
Department of Computer Science, Iowa State University
{mohnac, aditkulk, qli}@iastate.edu

Abstract

Recent studies have demonstrated that natural-language prompts can help to leverage the knowledge learned by pre-trained language models for the binary sentence-level sentiment classification task. Specifically, these methods utilize few-shot learning settings to fine-tune the sentiment classification model using manual or automatically generated prompts. However, the performance of these methods is sensitive to the perturbations of the utilized prompts. Furthermore, these methods depend on a few labeled instances for automatic prompt generation and prompt ranking. This study aims to find high-quality prompts for the given

user's intuition of the task (Schick and Schütze, 2021; Gao et al., 2021). Humans can easily write prompts, but the manual prompts are likely to be suboptimal since the language models may understand the instruction differently from humans. Prior studies have also shown that the performance of the language models is sensitive to the choice of prompts. For example, (Gao et al., 2021; Jiang et al., 2020) have shown that the performance is sensitive to the choice of certain words in the prompts and the position of the prompts. Due to the sensitivity and the potential misunderstanding of the instruction, manual prompts tend to suffer from poor performance under zero-shot settings.

Challenges & Main ideas

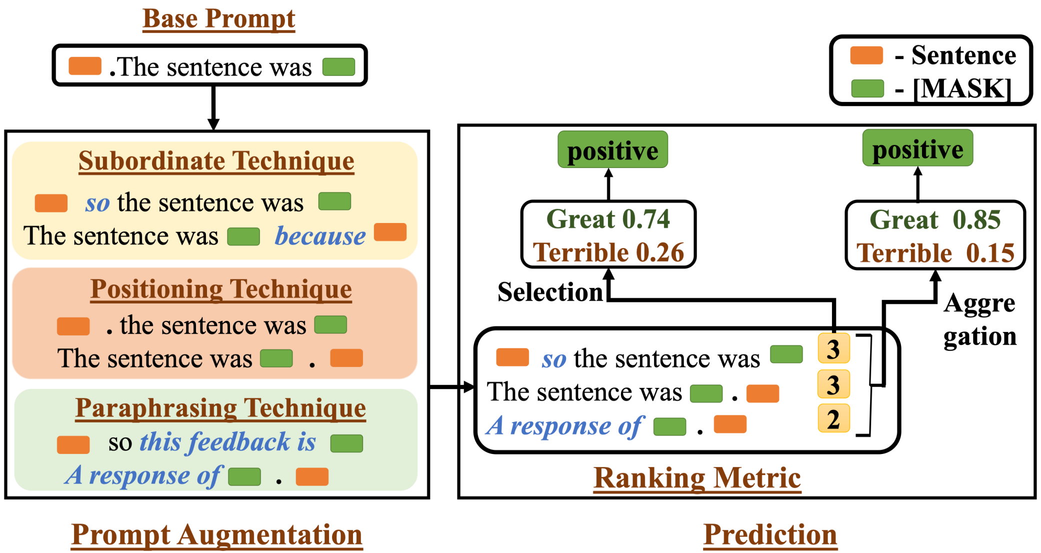
1. **Manually** prompts **sensitive to perturbation** [1, 2]
   → **Automatically generate** high-quality prompts
2. **Zero-shot setting**
3. **Prompt Generation** → **Ranking** → **Selection**
   - **Positioning, Subordination, Paraphrasing**
   - **Ranking metric:** sensitive to keyword change
4. **Task**: binary sentiment classification

Credits: [1] Gao *et al.*, Making Pre-trained Language Models Better Few-shot Learners, In ACL'21.
[2] Jiang *et al.*, How Can We Know What Language Models Know?, In TACL'20.

Part 3 – Prompt Perturbation Selected Works

Objectives

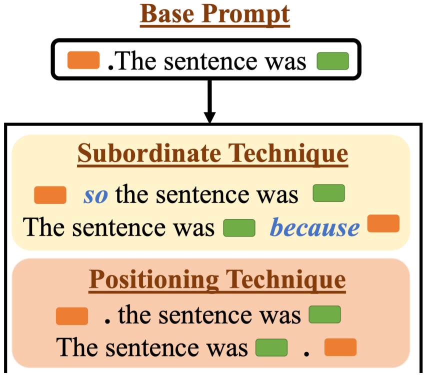
- ***Prompt Generation***
  - **Positioning** Technique
  - **Subordinate** Technique
  - **Paraphrasing** Technique
- ***Prompt Ranking***
  - **Zero-shot** Setting
- ***Prompt Selection***
  - **Prompt Selection** and **Aggregation**



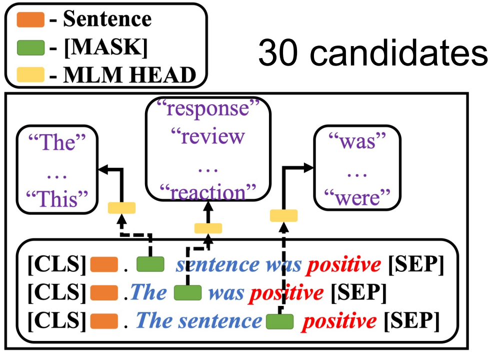
Credits: Chakraborty *et al.*, Zero-shot Approach to Overcome Perturbation Sensitivity of Prompts, In ACL'23.

Part 3 – Prompt Perturbation Selected Works

Prompt Augmentation – "[X]. The sentence was [Y]"



- **Positioning Technique**
  - ➢ Places the prompt either before or after the given sentence
  - ➢ "The sentence was [X]. [Y]"
- **Subordinate Technique**
  - ➢ Uses subordinate conjunctions like "because" and "so" to join the prompt and the sentence
  - ➢ "[X] so the sentence was [Y]" or "The sentence was [Y] because [X]"
- **Paraphrasing Technique**
  - ➢ Synonym Replacement (SR) to the base prompt B_p
  - ➢ Pre-trained MLM model $\mathcal{L}$ with a randomly selected sentence [X]
  - ➢ Mask the replaceable tokens from the base prompt one at a time



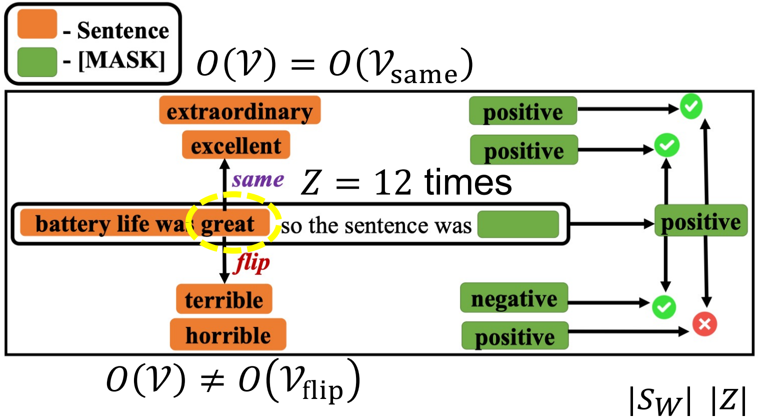
Credits: Chakraborty *et al.*, Zero-shot Approach to Overcome Perturbation Sensitivity of Prompts, In ACL'23.

Part 3 – Prompt Perturbation Selected Works

Prompt Ranking – under zero-shot setting

- **Zero-shot Setting**
  - High-quality prompt P (with number S_W) should be more sensitive to changing certain keywords $\mathcal{V}$
  - Key token $\mathcal{V}$ flips ⇒ Predicted label $\mathcal{Y}$ flips
  - Mapping token $\mathcal{V}$ [“**great**” → “**positive**”]
  - Use Wordnet [2] to obtain synonyms
  - Zero-one scoring function

$$\lambda_{s_{in}} = \begin{cases} 1, & \text{if } O(\mathcal{V}) = O(\mathcal{V}_{\text{same}}) \text{ or } O(\mathcal{V}) \neq O(\mathcal{V}_{\text{flip}}); \\ 0, & \text{otherwise.} \end{cases}$$



$$\text{Score}(P) = \sum_{i=1}^{|S_W|} \sum_{j=1}^{|Z|} \lambda_{s_{ij}}.$$

Credits:

[1] Chakraborty *et al.*, Zero-shot Approach to Overcome Perturbation Sensitivity of Prompts, In ACL’23.
[2] Miller *et al.*, WordNet: A Lexical Database for English, In Communications of the ACM’1995.

Part 3 – Prompt Perturbation Selected Works

Prompt Selection and Aggregation

- **Prompt Selection**
  - Given the sentence and prompt, predict $[\mathrm{MASK}]$ and select the highest probability

$$p(y|s_{\mathrm{in}}) = p([\mathrm{MASK}] \mid s_{\mathrm{in}}, P).$$

- **Prompt Aggregation**
  - Aggregate top-k ranked prompts

$$\mathrm{Score}(P_i) = \sum_{j=1}^{|Z|} \lambda_{s_j},$$

$$p(y) = \frac{\sum_{i=1}^{k} \mathrm{Score}(P_i) \times p_i(y)}{\sum_{i=1}^{k} \mathrm{Score}(P_i)}.$$

Credits: Chakraborty *et al.*, Zero-shot Approach to Overcome Perturbation Sensitivity of Prompts, In ACL’23.

Part 3 – Prompt Perturbation Selected Works

Data – *binary sentence-level sentiment classification datasets*

- **Stanford Sentiment Treebank v2** (SST-2) [2]: predicting Sentiment from longer Movie Reviews
- **MR Movie Reviews** (MR) [3]: overall sentiment polarity (positive or negative) or subjective rating (two and a half stars) and sentences with respect to their subjectivity status (subjective or objective) or polarity.
- **Customer Review** (CR) [4]: customer review of products

Models

- BERT-base and BERT-large models [5]

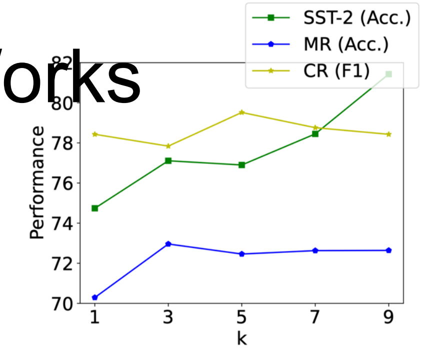
Datasets	**SST-2**		**MR**		**CR**	
	Pos	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**
Train	3610	3310	4331	4331	1407	368
Dev	444	428	0	0	0	0
Test	909	912	1000	1000	1000	1000
Total	4963	4650	5331	5331	2407	1368

Credits:
[1] Chakraborty *et al.*, Zero-shot Approach to Overcome Perturbation Sensitivity of Prompts, In ACL'23.
[2] Socher *et al.*, Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, In EMNLP'13.
[3] Pang *et al.*, Thumbs up? sentiment classification using machine learning techniques, In EMNLP'02.
[4] Hu *et al.*, Mining and summarizing customer reviews, In KDD'04.
[5] Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, In NAACL'19.

Part 3 – Prompt Perturbation Selected Works

Results

1. ⋆ base prompt **2.** aggregation strategy **3.** LM-BFF



manual + few-shot

fine-tuning

pretraining hard prompts by adding soft prompts

"<sentence>. **It** was [MASK]"

"<sentence>. **The sentence** was [MASK]"

Method	Prompt	BERT base						BERT large					
		SST-2		MR		CR		SST-2		MR		CR	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
LM-BFF	Automatic	58.46	62.24	57.94	62.81	71.35	69.66	52.69	59.33	57.3	63.69	70.55	69.11
UPT	Automatic	57.46	61.79	62.65	66.78	75.09	73.53	53.82	61.08	65.2	69.69	72.62	71.4
LM-BFF	Manual	62.3	65.75	58.18	62.16	74.9	72.81	61.15	65.41	57.88	62.64	72.59	70.85
PPT	Manual	52.53	56.93	50.5	53.41	64.03	61.02	52.29	57.68	50.5	56.0	63.9	62.21
Base Prompt†	Manual	62.3	65.75	58.18	62.16	74.9	72.81	61.15	65.41	57.88	62.64	72.59	70.85
Base Prompt⋆	Manual	63.22	63.15	59.97	60.25	69.04	64.29	54.12	58.6	54.43	57.12	56.59	62.14
ZS-SC (Top-1)†	Automatic	67.48	67.52	58.93	62.07	73.36	70.16	74.13	75.66	69.84	71.75	73.12	70.65
ZS-SC (Top-3)†	Automatic	67.12	68.22	60.15	60.14	71.19	68.23	67.58	70.65	64.15	67.91	70.05	67.82
ZS-SC (Top-5)†	Automatic	67.99	68.94	61.19	62.92	71.51	69.32	66.55	70.09	63.47	67.76	69.41	67.32
ZS-SC (Top-1)⋆	Automatic	**72.18**	**72.36**	**68.24**	**68.26**	75.09	72.1	74.74	74.71	70.29	70.36	80.47	78.43
ZS-SC (Top-3)⋆	Automatic	71.92	72.01	67.88	67.89	76.82	74.43	**77.11**	**77.58**	**72.96**	**73.54**	79.17	77.84
ZS-SC (Top-5)⋆	Automatic	71.5	71.46	66.74	66.88	**77.26**	**74.52**	76.9	77.54	72.46	73.43	**81.45**	**79.52**

Credits: Chakraborty *et al.*, Zero-shot Approach to Overcome Perturbation Sensitivity of Prompts, In ACL'23.

Part 3 – Prompt Perturbation Selected Works

Results

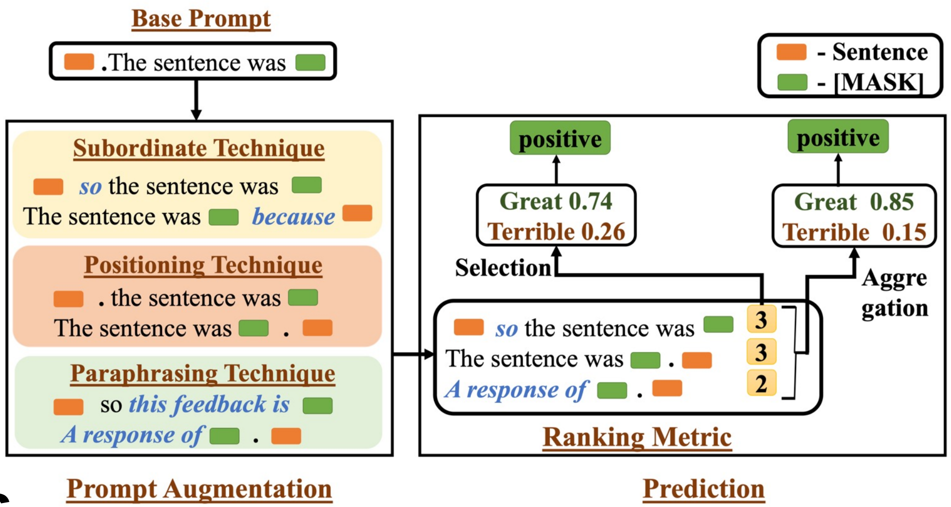
Top-ranked prompts

Dataset	BERT large	BERT base
SST-2	The sentence sounded [MASK] because <sentence> . Every sentence was [MASK] . <sentence> . <sentence> . Every sentence was [MASK] . The result was [MASK] . <sentence> . Each sentence was [MASK] . <sentence> .	<sentence>. Every sentence was [MASK] . Every sentence was [MASK]. <sentence> . Each sentence was [MASK] . <sentence> . <sentence>. Each sentence was [MASK] . <sentence> so every sentence was [MASK] .
MR	The sentence sounded [MASK] because <sentence> . The sentence seemed [MASK] because <sentence> . The result was positive . <sentence> . Every sentence was [MASK] because <sentence> . Every sentence was [MASK] . <sentence> .	<sentence>. Every sentence was [MASK] . Every sentence was [MASK]. <sentence> . Each sentence was [MASK] . <sentence> . <sentence> . Each sentence was [MASK] . <sentence> so the sentence sounded [MASK] .
CR	The sentence sounded [MASK] because <sentence> . The sentence sounded [MASK] . <sentence> . <sentence> . The sentence sounded [MASK] . Every sentence was [MASK] . <sentence> . The answer was [MASK] . <sentence> .	The sentence sounded [MASK] . <sentence> . <sentence> . The sentence sounded [MASK] . Every sentence was [MASK] . <sentence> . <sentence> . Every sentence was [MASK] . This sentence was [MASK] . <sentence> .

Dataset	LM-BFF	PPT	UPT
SST-2	<sentence>. A [MASK] one. <sentence>. A [MASK] piece. <sentence>. All in all [MASK].	<sentence>. [MASK].	<sentence>. It was [MASK]. <sentence>. I thought it was [MASK]. <sentence>. It is [MASK]. <sentence>. The review is [MASK]. <sentence>. A [MASK] one.
MR	It was [MASK] ! <sentence>. <sentence>. It's [MASK]. <sentence> A [MASK] piece of work.	<sentence>. [MASK].	<sentence>. A [MASK] piece of work. <sentence>. It is [MASK]. <sentence>. The film is [MASK]. <sentence>. A really [MASK] movie.
CR	<sentence>. It's [MASK] ! <sentence>. The quality is [MASK]. <sentence>. That is [MASK].	<sentence>. [MASK].	<sentence>. It was [MASK]. <sentence>. It looks [MASK]. <sentence>. It is [MASK]. <sentence>. The quality is [MASK]. <sentence>. I thought it was [MASK].

Credits: Chakraborty *et al.*, Zero-shot Approach to Overcome Perturbation Sensitivity of Prompts, In ACL'23.

Part 3 – Prompt Perturbation Selected Works



Limitations

- ***Scenarios***: **limited area of output**, *e.g.*, positive or negative
- ***Subordinate***: **because-so causality**
- ***Prediction***: **single-token objects**
- ***Ranking***: Need **mapping token**

Credits: Chakraborty *et al.*, Zero-shot Approach to Overcome Perturbation Sensitivity of Prompts, In ACL'23.

Part 3 – Prompt Perturbation Selected Works

Summary

Year	Author	Institution	Title	Scenario	Metrics	Method	Model	Result
2020 TACL	**Zhengbao Jiang**, Frank F. Xu, Jun Araki, and **Graham Neubig**	CMU, Bosch Research	How Can We Know What Language Models Know?	Relation triples	Micro-averaged Accuracy; Macro-averaged Accuracy	Mining Paraphrasing	BERT-base; BERT-large	Mine+Man: 43.9% (Micro) and 30.7% (Macro) on LAMA T-REx
2023 ACL	Mohna Chakraborty, Adithya Kulkarni, and Qi Li	Iowa State University	Zero-shot Approach to Overcome Perturbation Sensitivity of Prompts	Limited area of output (positive or negative)	Accuracy; Macro F1 Score	Subordinate Positioning Paraphrasing	BERT-base; BERT-large	Accuracy: 77.11% (SST-2) 72.96% (MR) 81.45% (CR)
2023 EACL	Yoichi Ishibashi, Danushka Bollegala, *et al.*	Nara Institute of Science and Technology, ULiverpool	Evaluating the Robustness of Discrete Prompts	Evaluation of prompt perturbation	Accuracy; Rate of Degradation (RoD)	Token Reordering Deletion; Adversarial Perturbations	AutoPrompt; Manually-written Prompts (MP)	-

Part 4 – Robustness Problem Formulation

- **Foundational Robustness**:
  - Evaluation and enhancement of (and sometimes certifiable) model correctness against natural and adversarial data shifts → A foundation of trustworthy AI
- **Robustness Category**:
  - **Adversarial Robustness** (worst-case performance)
    $\boldsymbol{x}'$ similar to $\boldsymbol{x}$, and $\boldsymbol{\delta}$ is small perturbations.
    Ideally, $f_\theta(\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{\delta}) = f_\theta(\boldsymbol{x})$.
  - **Out-of-distribution (OOD) generalization** (domain shifts)
    $\boldsymbol{x}{\sim}D,\ \boldsymbol{x}'{\sim}D'$, where D' is the shifted version of D.
    Ideally, $f_\theta(\boldsymbol{x}') = f_\theta(\boldsymbol{x})$.
  - **Out-of-distribution detection** (unknowns)
    $\boldsymbol{x}{\sim}D,\ \boldsymbol{x}'{\sim}D'$, where D' is a dissimilar or new domain compared with D.
    Ideally, $f_\theta(\boldsymbol{x}') = \text{“Unknown”}$.

Credits: Pin-Yu Chen and Sijia Liu, Foundational Robustness of Foundation Models Tutorial, In NeurIPS’22.

Part 4 – Robustness Problem Formulation

- **Empirical Adversarial Robustness**:

The model $f(\cdot)$ is robust by optimizing the empirical adversarial risk:

$$\min_{\theta} \mathbb{E}_{(\boldsymbol{x},\, \boldsymbol{y})} \left[\max_{\boldsymbol{\delta} \in \Delta} \mathcal{L}(f_{\theta}(\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{\delta}), \boldsymbol{y})\right].$$

- Robust optimization (min-max) formulation of adversarial learning
- Δ: a neighborhood (allowable subset of perturbations) of $\boldsymbol{x}$
- $\boldsymbol{\delta} \in \Delta = \{\boldsymbol{\delta}\colon \|\boldsymbol{\delta}\|_{\infty} = \max_i |\boldsymbol{\delta}_i| \leq \epsilon\}$, and $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{\delta}$ is the adversarial example
- $\mathcal{L}$: negative cross entropy of $f_{\theta}(\boldsymbol{x})$ and $\boldsymbol{y}$
- Provide robustness within an ϵ-bounded *threat model* for an ℓ_p or ℓ_{∞} norm

Credits: Pin-Yu Chen and Sijia Liu, Foundational Robustness of Foundation Models Tutorial, In NeurIPS'22.

Part 4 – Robustness Problem Formulation

- **Certified Adversarial Robustness**:

  The model $f(\cdot)$ is certified robust if it satisfies the following condition for $\forall \boldsymbol{x}$:

$$f(\boldsymbol{x}') = f(\boldsymbol{x}) = y,$$
$$\|\boldsymbol{x}' - \boldsymbol{x}\|_0 = \sum_{i=1}^{L} \mathbb{I}(\boldsymbol{x}'_i \neq \boldsymbol{x}_i) \leq dL.$$

  - $\boldsymbol{x} = [x_1, x_2, \ldots, x_L]$: input to the LLM $f(\cdot)$
  - $\|\boldsymbol{x}' - \boldsymbol{x}\|_0$: Hamming Distance
  - $\mathbb{I}(\cdot)$: Indicator Function
  - d: perturbation scale; dL: neighborhood R (certified range)

- **Problem**: **(1)** Same length sequence **(2)** Never consider semantics change

Credits: Zhang *et al.*, Certified Robustness for Large Language Models with Self-Denoising, In arXiv'23.

Part 5 – Robustness Evaluation

- **Rate of Degradation (RoD)** [1, 2] / **MultiModal Impact score (MMI)** [3]:
- The decrease in accuracy of the target task due to the perturbations added to the prompt.
- A smaller RoD indicates a more robust model against perturbations

$$\mathrm{RoD} = \frac{\mathrm{avgacc}_x - \mathrm{avgacc}_{x^*}}{\mathrm{avgacc}_x} = 1 - \frac{\mathrm{avgacc}_{x^*}}{\mathrm{avgacc}_x},$$

- where x^* is the perturbed version of the original prompt x, and avgacc_x and avgacc_{x^*} are the averaged accuracies over M prompts

Credits:

[1] Meyers *et al.*, Signal Processing on PV Time-series Data: Robust Degradation Analysis Without Physical Models, In IEEE J-PV'19.
[2] Ishibashi *et al.*, Evaluating the Robustness of Discrete Prompts, In EACL'23.
[3] Qiu *et al.*, Are Multimodal Models Robust to Image and Text Perturbations?, In arXiv'23.

Thank you very much for your attention!

Dependable Computing Laboratory,
Department of Electrical and Computer Engineering,
Boston University

BOSTON UNIVERSITY