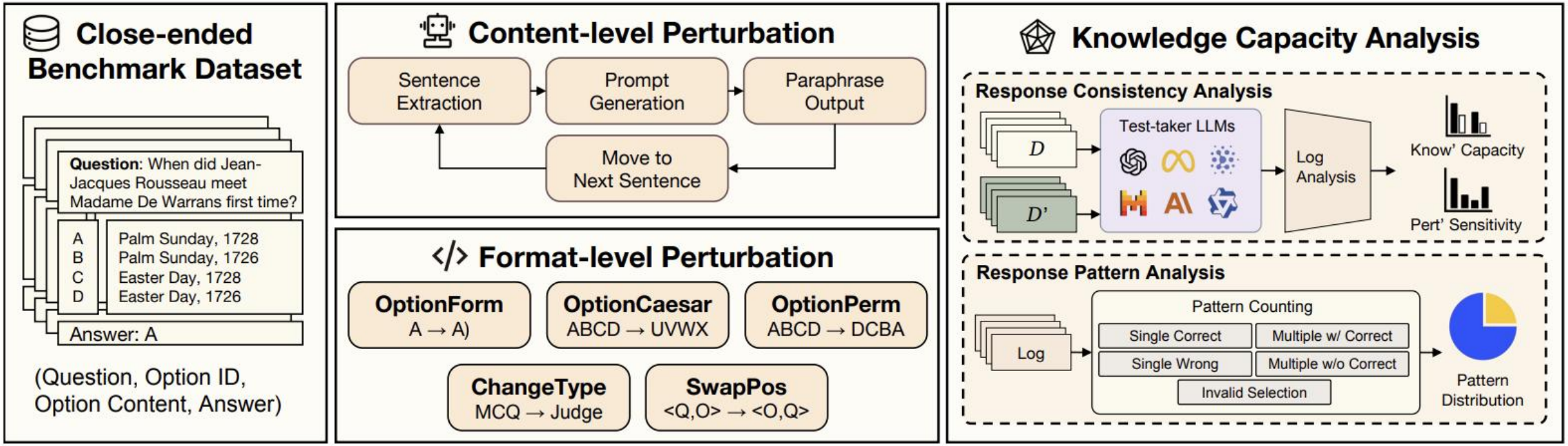
# PertEval: Unveiling Real Knowledge Capacity of LLMs via Knowledge-Invariant Perturbations

**Jiatong Li**[1], Renjun Hu[2], Kunzhe Huang[2], Yan Zhuang[1],
Qi Liu[1], Mengxiao Zhu[1], Xing Shi[2], Wei Lin[2]

[1]University of S&T of China, [2]Alibaba Cloud Computing

# Contents

# 1.1. Research Overview




- **Research Task:** Multiple Choice Question-based LLM Knowledge Eval
- **Challenge:** Data Contamination; Limited Test Scenarios → **Untruthful Eval**
- **Solution:** A Knowledge-invariant Perturbation-based Eval Toolkit

# 1.2. Contributions

- We propose **PertEval** to unveil the real knowledge capacity of LLMs, marking a significant step towards more **trustworthy** LLM evaluation.
- We re-evaluate six LLMs using PertEval. Evaluation results reveal **overestimated** performance of LLMs and their **uncertainty** to specious knowledge.
- We demonstrate the **vulnerability** of LLMs to different perturbation strategies in PertEval and provide insights for the **refinement** of knowledge capacity.

# 2. Architecture of PertEval

Table 7: An example of knowledge-invariant paraphrasing of a test question. Texts surrounded by angular brackets are invisible in question prompts input to the LLM test-taker.

| Original | Knowledge-invariant Paraphrasing |
|---|---|
| <# **Context & Condition**> Let $T : R^2 \to R^2$ be the linear transformation that maps the point (1, 2) to (2, 3) and the point (-1, 2) to (2, -3). <# **Goal**> Then $T$ maps the point (2, 1) to | <# **Context & Condition**> Let $T$ be the linear transformation from $R^2$ to $R^2$ such that $T$ maps (1, 2) to (2, 3) and (-1, 2) to (2, -3). <# **Goal**> Then, the linear transformation $T$ will map the point (2, 1) to |

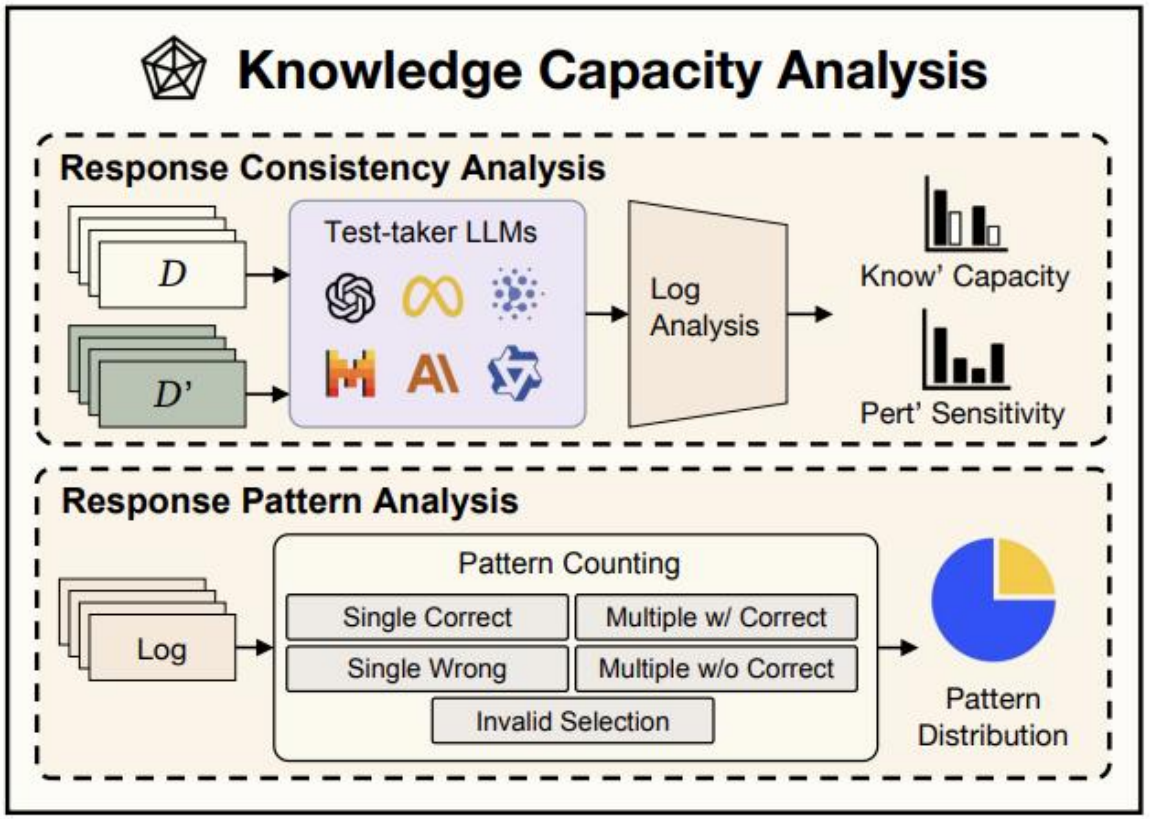
Table 8: Examples of format-level knowledge-invariant perturbations. Texts surrounded by angular brackets are invisible in question prompts input to the LLM test-taker.

| Perturbation | Original case | Perturbed case |
|---|---|---|
| OptionPerm | <# **Options**> A $x = 1$; B $x = 2$; C $x = 3$; D $x = 4$ | <# **Options**> A $x = 4$; B $x = 3$; C $x = 2$; D $x = 1$ |
| OptionForm | <# **Options**> A $x = 1$; B $x = 2$; C $x = 3$; D $x = 4$ | <# **Options**> A) $x = 1$; B) $x = 2$; C) $x = 3$; D) $x = 4$ |
| OptionCaesar | <# **Options**> A $x = 1$; B $x = 2$; C $x = 3$; D $x = 4$ | <# **Options**> U $x = 1$; V $x = 2$; W $x = 3$; X $x = 4$ |
| ChangeType | <# **Prompt**> Please select correct option(s) given the following question: | <# **Prompt**> Please judge whether each of the options is correct given the following question: |
| SwapPos | <# **Prompt**> Please select correct option(s) given the following question: <# **Question**> The solution of the equation $2x + 1 = 3$ is <# **Options**> A $x = 1$; B $x = 2$; C $x = 3$; D $x = 4$ | <# **Prompt**> Please select correct option(s) given the following question: <# **Options**> A $x = 1$; B $x = 2$; C $x = 3$; D $x = 4$ <# **Question**> The solution of the equation $2x + 1 = 3$ is |

Examples of knowledge-invariant perturbation.

- **Content-level Perturbation**
  - Knowledge-invariant paraphrasing
- **Format-level Perturbation**
  - Five different strategies
- **Knowledge Capacity Analysis**
  - Response Consistency Analysis
  - Response Pattern Analysis

# 2. Architecture of PertEval




- **Content-level Perturbation**
  - Knowledge-invariant paraphrasing
- **Format-level Perturbation**
  - Five different strategies
- **<u>Knowledge Capacity Analysis</u>**
  - Response Consistency Analysis
  - Response Pattern Analysis

# 3. Knowledge Invariance Verification

Table 2: **Knowledge invariance scores↑ rated by human scorers**. Four independent scores from different human scorer groups are presented in ascending order for each cell.

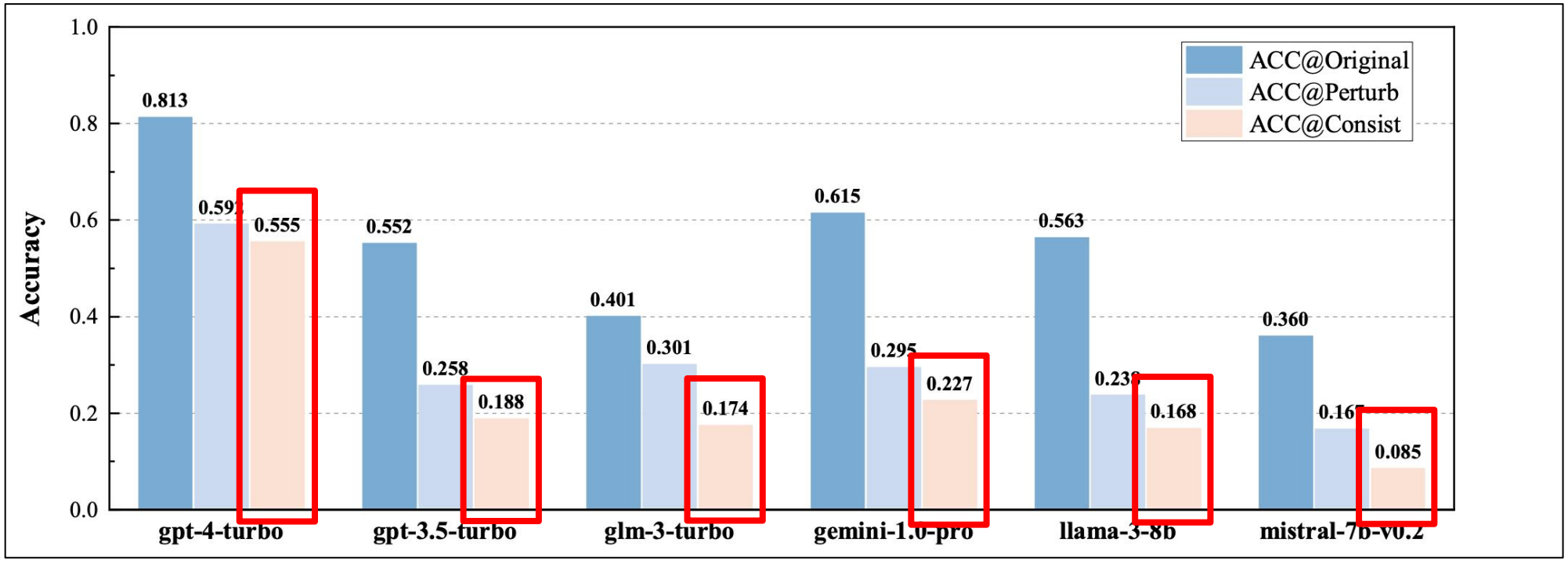
| Method | C-Math | W-History | P-Psychology | P-Medicine |
|---|---|---|---|---|
| PromptAttack | 2.3/2.4/3.0/3.9 | 2.2/2.2/2.3/2.8 | 1.3/2.4/2.8/2.8 | 1.6/2.5/3.5/3.6 |
| **PertEval (ours)** | **3.6/3.8/3.9/3.9** | **3.7/4.1/4.1/4.3** | **4.3/4.4/4.5/4.7** | **4.2/4.3/4.4/4.6** |

Table 3: **Knowledge invariance scores↑ rated by superior LLMs**. Values (a/b/c) in each cell denotes the average knowledge invariance score rated by gpt-4-turbo, claude-3.5-sonnet, and llama-3.1-405b, respectively.

| Method | C-Math | W-History | P-Psychology | P-Medicine |
|---|---|---|---|---|
| PromptAttack | 3.2/3.6/3.6 | 3.2/3.3/3.7 | 3.9/3.9/3.7 | 4.1/4.3/4.2 |
| **PertEval (ours)** | **3.8/3.9/4.0** | **4.0/4.2/4.0** | **4.0/4.4/4.0** | **4.1/4.4/4.0** |

- **Verification Method:** Human-based & LLM-based scoring (min: 1; max: 5)
- **Experiment Dataset:** A subset of MMLU covering all 4 major topics
- **Findings:** PertEval obtains high KI scores ($\geq 3.6$ for C-Math; mostly $\geq 4.0$ for others)
- **Conclusion:** PertEval can indeed generate knowledge-invariant perturbed datasets

# 4.1. Response Consistency Analysis




- **Evaluation Metric:** ACC@Consist (Ratio of questions that are correctly answered on both the original & perturbed datasets)
- **Finding**: Overestimated knowledge capacity on the orignal dataset

# 4.1. Response Consistency Analysis

Table 5: **Macro PDR** ↑ and **hypothesis test results of Micro PDR** of LLMs w.r.t. perturbation.

| Model\Strategy | KnInvPara | OptionPerm | OptionForm | OptionCaesar | ChangeType | SwapPos | AVG |
|---|---|---|---|---|---|---|---|
| **gpt-4-turbo** | -0.0660** | -0.0208** | -0.0136 | -0.0294** | -0.0210 | -0.1117** | -0.0438 |
| **gpt-3.5-turbo** | -0.0275** | -0.0042 | -0.1767** | -0.0396** | -0.1736** | -0.1943** | -0.1027 |
| **gemini-1.0-pro** | -0.0558** | +0.0121 | +0.0125 | +0.0030 | -0.1310** | -0.1532** | -0.0521 |
| **glm-3-turbo** | -0.0370** | -0.0190 | -0.1397** | -0.0118 | +0.0522 | -0.2142** | -0.0616 |
| **mistral-7b-v0.2** | -0.0264 | -0.0200 | -0.2789** | +0.0793 | -0.0844** | -0.1275** | -0.0763 |
| **llama-3-8b** | -0.0336** | -0.0091 | -0.0939** | -0.0368** | -0.2920** | -0.1814** | -0.1078 |
| **AVG** | -0.0411 | -0.0102 | -0.1151 | -0.0059 | -0.1083 | -0.1637 | |

**: The *Micro* PDR is significantly negative in the Wilcoxon signed-rank test ($\alpha = 0.01$).

- **Evaluation Metric**: Performance Drop Rate (PDR); Recall of Performance (ROP)
- **Finding**: 1. The sensitivity of LLMs to perturbations differs a lot; 2. The effect of perturbations differ a lot; 3. All LLMs are sensitive to SwapPos, the global order change.

# 4.2. Response Pattern Analysis

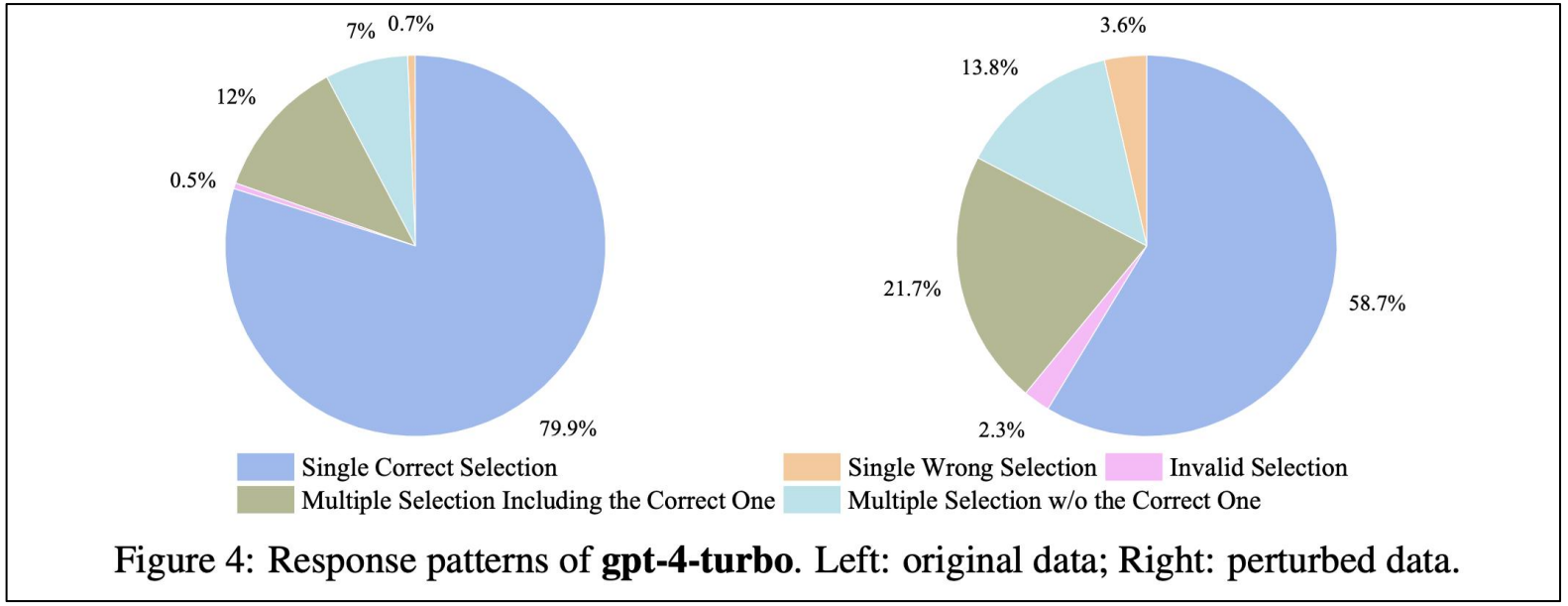



Figure 4: Response patterns of **gpt-4-turbo**. Left: original data; Right: perturbed data.

- **Evaluation Method**: Count & Visualize & Compare response patterns on the original & perturbed datasets.
- **Finding**: For most LLMs, the ratio of **multiple selection including the correct one** significantly increases on the perturbed dataset, indicating their uncertainty to incorrect knowledge.

# 5. Empowering LLMs' Capacity

| Strategy | fine-tune | C-Math | W-History | P-Psychology | P-Medicine | $AVG_{macro}$ | $AVG_{micro}$ |
|---|---|---|---|---|---|---|---|
| ChangeType | None | **-0.2300** | **-0.3418** | **-0.2467** | **-0.3493** | **-0.2920** | **-0.1998** |
| | F(CT) | -0.0700 | +0.0759 | +0.0196 | +0.0257 | +0.0128 | -0.0868 |
| | F(CT+KP) | -0.0500 | +0.0422 | +0.0082 | +0.0074 | +0.0020 | +0.0019 |
| SwapPos | None | **-0.0700** | **-0.2110** | **-0.1944** | **-0.2500** | **-0.1814** | **-0.2867** |
| | F(SP) | +0.0100 | -0.0675 | -0.1095 | -0.0882 | -0.0638 | +0.0246 |
| | F(SP+KP) | -0.0300 | -0.1350 | -0.1029 | -0.1176 | -0.0964 | -0.1065 |
| KnInvPara | None | +0.0200 | **-0.0802** | -0.0163 | **-0.0478** | -0.0311 | -0.0328 |
| | F(KP) | **-0.0400** | -0.0253 | -0.0212 | 0.0000 | -0.0216 | -0.0188 |
| | F(CT+KP) | **-0.0400** | -0.0549 | **-0.0343** | -0.0368 | **-0.0415** | **-0.0393** |
| | F(SP+KP) | -0.0300 | -0.0675 | -0.0212 | -0.0184 | -0.0343 | -0.0303 |

- **Method**: Supervised fine-tuning llama-3-8b-instruct using perturbed W-History & P-Medicine
- **Finding**:
  - **Stimulation Phenomenon**: **For format-level perturbations**, only fine-tuning the model with a subset of perturbed data can significantly improve its overall performance stability in all perturbed data.

# 5. Empowering LLMs' Capacity

| Strategy | fine-tune | C-Math | W-History | P-Psychology | P-Medicine | $AVG_{macro}$ | $AVG_{micro}$ |
|---|---|---|---|---|---|---|---|
| ChangeType | None | **-0.2300** | **-0.3418** | **-0.2467** | **-0.3493** | **-0.2920** | **-0.1998** |
| | F(CT) | -0.0700 | +0.0759 | +0.0196 | +0.0257 | +0.0128 | -0.0868 |
| | F(CT+KP) | -0.0500 | +0.0422 | +0.0082 | +0.0074 | +0.0020 | +0.0019 |
| SwapPos | None | **-0.0700** | **-0.2110** | **-0.1944** | **-0.2500** | **-0.1814** | **-0.2867** |
| | F(SP) | +0.0100 | -0.0675 | -0.1095 | -0.0882 | -0.0638 | +0.0246 |
| | F(SP+KP) | -0.0300 | -0.1350 | -0.1029 | -0.1176 | -0.0964 | -0.1065 |
| KnInvPara | None | +0.0200 | **-0.0802** | -0.0163 | **-0.0478** | -0.0311 | -0.0328 |
| | F(KP) | **-0.0400** | -0.0253 | -0.0212 | 0.0000 | -0.0216 | -0.0188 |
| | F(CT+KP) | **-0.0400** | -0.0549 | **-0.0343** | -0.0368 | **-0.0415** | **-0.0393** |
| | F(SP+KP) | -0.0300 | -0.0675 | -0.0212 | -0.0184 | -0.0343 | -0.0303 |

- **Method**: Supervised fine-tuning llama-3-8b-instruct using perturbed W-History & P-Medicine
- **Finding**:
  - **Lack of Transferability**: **For content-level perturbations**, SFT on a subset of the perturbed datasets cannot significantly improve its performance on other perturbed subsets (subjects).

# 6. Conclusion

- **One** trustworthy evaluation toolkit - PertEval
- **Two** response analysis methods - consistency & pattern analyses
- **Three** evaluation metrics - ACC@Consist, PDR, ROP
- **Four** significant experiments
- **Five** format-level perturbations
- **Six** perturbations in total for PertEval

# PertEval: Unveiling Real Knowledge Capacity of LLMs via Knowledge-Invariant Perturbations

Arxiv: https://arxiv.org/abs/2405.19740

Code: https://github.com/aigc-apps/PertEval


NEURAL INFORMATION
PROCESSING SYSTEMS


Alibaba Cloud


1958

# Thank you!