# InstructEval: Systematic Evaluation of Instruction Selection Methods

Paper Code

Anirudh Ajith*, Chris Pan*, Mengzhou Xia, Ameet Deshpande, Karthik Narasimhan

## Motivation

### How do we optimize in-context learning performance?

Is this review positive or negative?

**Review:** Whoever wrote the screenplay for this movie obviously never consulted...
**Sentiment:** Negative

**Review:** The story centers around Barry McKenzie who must go to England...
**Sentiment:** Positive

**Review:** This film is just plain horrible. John Ritter doing pratt falls, 75% of the actors...
**Sentiment:** Negative

**Review:** BLACK WATER has to be one of the best Australian movies I've seen in many...
**Sentiment:**

Which demonstrations?
explored! 🔍 ✅

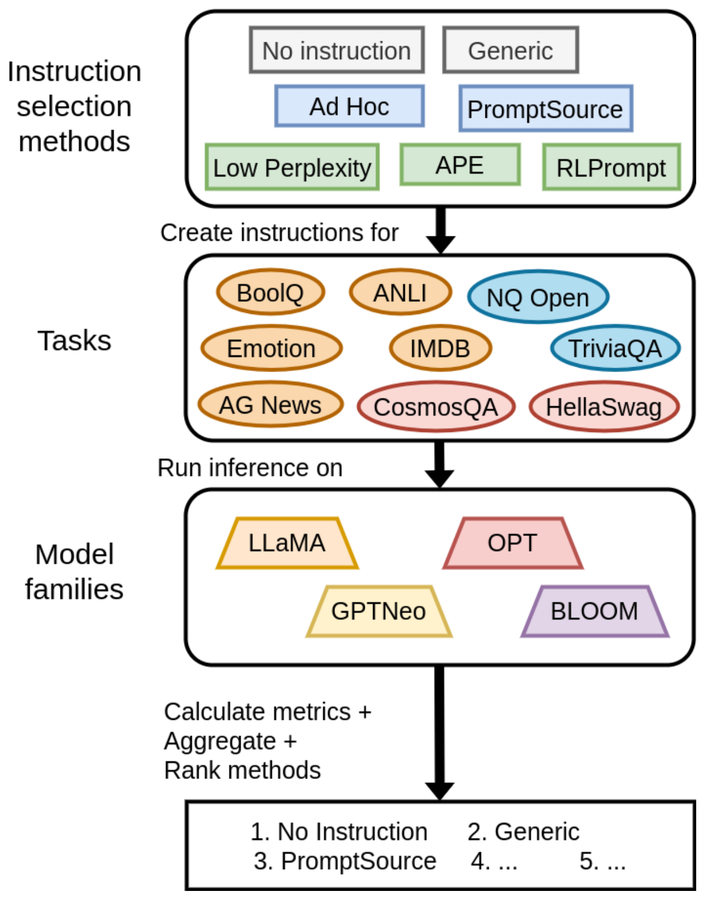
Which instructions?
underexplored! ❓ ⚠️

Existing instruction selection works [1, 2, 3]

- evaluate on tasks and models with little mutual intersection.
- focus on zero-shot accuracy.
- focus on classification tasks.

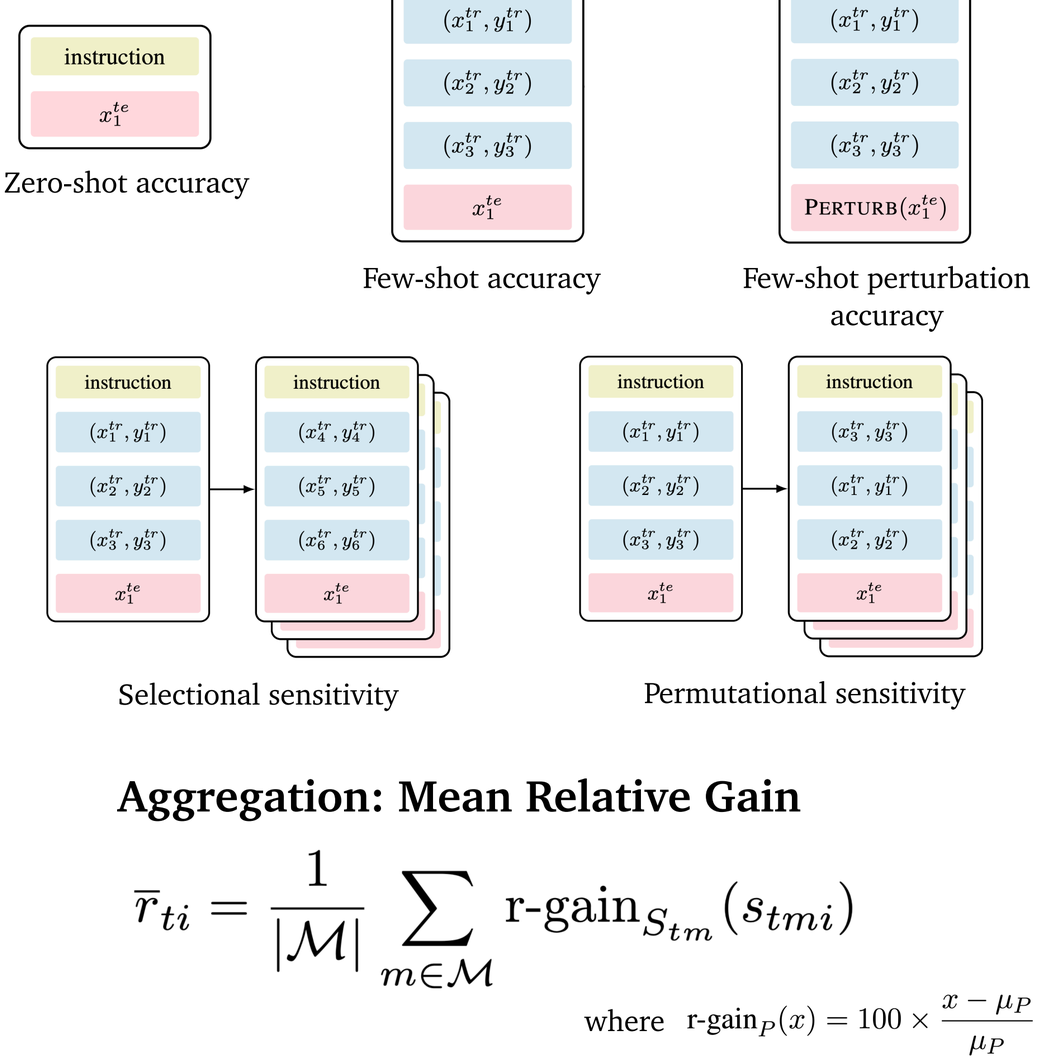
## InstructEval

### Holistic comparison of instruction selection methods!




- **9 tasks** spanning classification, multiple-choice question-answering and generation.
- **13 models** spanning 4 model families [1.1B - 13B].
- **5 metrics** for practical in-context learning.

### Metrics: Accuracy and Senstivity




### Aggregation: Mean Relative Gain

$$\overline{r}_{ti} = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \text{r-gain}_{S_{tm}}(s_{tmi})$$

$$\text{where} \quad \text{r-gain}_P(x) = 100 \times \frac{x - \mu_P}{\mu_P}.$$

## Results

| Method | AG News | ANLI | BoolQ | IMDB | Emotion | HellaSwag | CosmosQA | TriviaQA | NQ-Open | # wins |
|---|---|---|---|---|---|---|---|---|---|---|
| | CLS | | | | | MCQ | | GQA | | |
| **Zero-shot accuracy (mean relative gain) ↑** | | | | | | | | | | |
| Null Instruction | 2.26 | 1.07 | **2.48** | −3.52 | −5.30 | **2.54** | **5.94** | −3.08 | −25.67 | 3 |
| Generic Instruction | 3.55 | −0.39 | 0.03 | 1.69 | 2.39 | −0.13 | −1.67 | −1.52 | −5.99 | 0 |
| PromptSource | **5.81** | **1.38** | −0.65 | **4.34** | **5.13** | −1.54 | −3.42 | **17.02** | **22.15** | 6 |
| Ad hoc | −0.33 | 0.21 | 0.55 | 1.41 | 0.66 | −0.27 | −2.46 | −2.03 | 2.31 | 0 |
| Low Perplexity | −0.59 | 1.22 | 0.56 | 0.84 | −4.07 | −1.38 | −2.18 | −5.87 | 2.81 | 0 |
| APE | −15.63 | −3.86 | −1.07 | −1.77 | −0.26 | −1.06 | 0.00 | −4.70 | 4.39 | 0 |
| RLPrompt | 4.92 | 0.37 | −1.89 | −2.99 | 1.46 | 1.85 | 3.79 | – | – | 0 |
| **Few-shot accuracy (mean relative gain) ↑** | | | | | | | | | | |
| Null Instruction | 4.09 | −0.22 | **0.87** | −0.80 | **5.89** | 0.17 | **1.33** | **0.45** | −0.02 | 4 |
| Generic Instruction | **5.16** | −0.20 | −0.10 | 0.45 | 4.84 | 0.04 | −0.18 | 0.11 | 0.11 | 1 |
| PromptSource | 0.83 | 0.14 | −0.79 | 0.39 | −4.39 | −0.06 | −0.94 | −0.36 | **0.61** | 1 |
| Ad hoc | 2.18 | −0.10 | −0.05 | **0.60** | −5.63 | −0.21 | −0.59 | 0.09 | −0.49 | 1 |
| Low Perplexity | −1.96 | **0.31** | −0.40 | 0.20 | −6.79 | −0.23 | −0.61 | −0.06 | −0.02 | 1 |
| APE | −15.43 | 0.10 | 0.06 | −0.69 | 1.17 | 0.02 | 0.17 | −0.24 | −0.19 | 0 |
| RLPrompt | 5.13 | −0.02 | 0.40 | −0.14 | 4.90 | **0.27** | 0.81 | – | – | 1 |
| **Few-shot perturbation accuracy (mean relative gain) ↑** | | | | | | | | | | |
| Null Instruction | 4.09 | −0.08 | 0.11 | −0.27 | **5.98** | 0.11 | **1.10** | **0.81** | **1.28** | 4 |
| Generic Instruction | **5.15** | −0.18 | −0.16 | **0.56** | 4.23 | −0.02 | −0.02 | 0.08 | 0.10 | 2 |
| PromptSource | 1.14 | 0.27 | −0.02 | 0.33 | −3.92 | 0.06 | −0.53 | −0.65 | 0.04 | 0 |
| Ad hoc | 1.68 | 0.51 | −0.34 | 0.37 | −5.87 | −0.08 | −0.63 | −0.28 | −0.61 | 0 |
| Low Perplexity | −2.39 | **0.68** | −0.12 | −0.20 | −6.61 | −0.09 | −0.66 | −0.03 | −0.78 | 1 |
| APE | −14.32 | −1.20 | **0.28** | −0.82 | 1.26 | −0.13 | 0.21 | 0.06 | −0.03 | 1 |
| RLPrompt | 4.65 | −0.01 | 0.24 | 0.03 | 4.94 | **0.15** | 0.53 | – | – | 1 |

- Curated PromptSource [4] instruction dominate zero-shot.
- Task-agnostic instructions dominate few-shot settings.
- Automatic instruction selection methods outperformed by simple baselines!

| Method | AG News | ANLI | BoolQ | IMDB | Emotion | HellaSwag | CosmosQA | TriviaQA | NQ-Open | # wins |
|---|---|---|---|---|---|---|---|---|---|---|
| | CLS | | | | | MCQ | | GQA | | |
| **Selectional sensitivity (mean standard deviation) ↓** | | | | | | | | | | |
| Null Instruction | **6.69** | 2.45 | 4.73 | **5.28** | 6.97 | 2.46 | **8.10** | 2.59 | 2.28 | 3 |
| Generic Instruction | 6.87 | 2.50 | 4.76 | 5.40 | 6.97 | 2.48 | 8.16 | 2.61 | 2.26 | 0 |
| PromptSource | 6.73 | 2.26 | 4.85 | 5.37 | 6.43 | 2.43 | 8.26 | **2.59** | 2.28 | 1 |
| Ad hoc | 6.95 | 2.41 | **4.62** | 5.38 | 6.34 | 2.42 | 8.20 | 2.65 | 2.37 | 1 |
| Low Perplexity | 7.07 | **2.17** | 4.69 | 5.64 | **6.25** | 2.42 | 8.27 | 2.59 | 2.30 | 2 |
| APE | 7.44 | 2.98 | 4.63 | 5.70 | 6.67 | 2.43 | 8.16 | 2.65 | **2.21** | 1 |
| RLPrompt | 6.76 | 2.30 | 4.79 | 5.50 | 6.96 | **2.36** | 8.16 | – | – | 1 |
| **Permutational sensitivity (mean standard deviation) ↓** | | | | | | | | | | |
| Null Instruction | 6.02 | **1.99** | 3.82 | **4.14** | 5.48 | 1.12 | 1.87 | 1.52 | 1.28 | 2 |
| Generic Instruction | **6.01** | 2.19 | 3.89 | 4.56 | 5.49 | 1.15 | 1.68 | **1.33** | 1.22 | 2 |
| PromptSource | 6.06 | 2.15 | 3.61 | 4.69 | 4.30 | **1.07** | 1.67 | 1.47 | **1.17** | 2 |
| Ad hoc | 6.10 | 2.37 | 3.77 | 4.61 | 4.37 | 1.11 | 1.66 | 1.41 | 1.23 | 0 |
| Low Perplexity | 6.13 | 2.24 | **3.50** | 4.61 | **4.29** | 1.13 | 1.69 | 1.46 | 1.27 | 2 |
| APE | 6.14 | 2.36 | 3.69 | 4.84 | 5.08 | 1.10 | 1.78 | 1.41 | 1.21 | 0 |
| RLPrompt | 6.26 | 2.06 | 3.82 | 4.89 | 5.64 | 1.08 | **1.65** | – | – | 1 |

- All methods show similar sensitivity to selection and permutation of demonstrations.

## Takeaways

Existing automatic instruction selection methods
- do not generalize well to more models and tasks.
- may require extensive hyperparameter tuning.
- can be compuationally expensive.

Prompts that work well for one model/task may not transfer.
- Setting-specific search may be unavoidable.

Recommendations for practical scenarios:
- Use curated instructions (eg. PromptSource [4]) in zero-shot prompts.
- Don't use instructions in few-shot prompts.
- Use few-shot prompting whenever possible.

More systematic research towards automated instruction selection methods is needed. We release the InstructEval evaluation suite to aid in this research.

**References**:
**[1]** Gonen, H., Iyer, S., Blevins, T., Smith, N.A., & Zettlemoyer, L. (2022). Demystifying Prompts in Language Models via Perplexity Estimation. *ArXiv, abs/2212.04037*.
**[2]** Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., & Ba, J. (2022). Large Language Models Are Human-Level Prompt Engineers. *ArXiv, abs/2211.01910*
**[3]** Deng, M., Wang, J., Hsieh, C.-P., Wang, Y., Guo, H., Shu, T., Song, M., Xing, E., & Hu, Z. (2022). RLPrompt: Optimizing Discrete Text Prompts with Reinforcement Learning. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.
**[4]** Bach, S., Sanh, V., Yong, Z. X., Webson, A., Raffel, C., Nayak, N. V., Sharma, A., Kim, T., Bari, M. S., Fevry, T., Alyafeai, Z., Dey, M., Santilli, A., Sun, Z., Ben-david, S., Xu, C., Chhablani, G., Wang, H., Fries, J., … Rush, A. (2022). PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics.