**Motivation**

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

- **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.

- Task-agnostic instructions dominate few-shot settings.
- Automatic instruction selection methods outperformed by simple baselines!

**Results**

| Method | AG News | ANLI | BookReview | DUBSP | Emotion | FashionQ | GoogleNews | CommonsQA | TrenzQA | QQA
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Null Instruction</td>
<td>6.69</td>
<td>4.05</td>
<td>4.72</td>
<td>5.28</td>
<td>5.69</td>
<td>4.97</td>
<td>5.16</td>
<td>4.88</td>
<td>5.10</td>
<td>2.28</td>
</tr>
<tr>
<td>PromtSource</td>
<td>6.67</td>
<td>4.06</td>
<td>4.74</td>
<td>5.28</td>
<td>5.72</td>
<td>4.98</td>
<td>5.16</td>
<td>4.88</td>
<td>5.10</td>
<td>2.28</td>
</tr>
<tr>
<td>Ad-hoc</td>
<td>6.70</td>
<td>4.08</td>
<td>4.72</td>
<td>5.31</td>
<td>5.74</td>
<td>4.99</td>
<td>5.16</td>
<td>4.89</td>
<td>5.10</td>
<td>2.28</td>
</tr>
<tr>
<td>Low Perplexity</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>APE</td>
<td>6.77</td>
<td>4.11</td>
<td>4.74</td>
<td>5.29</td>
<td>5.75</td>
<td>5.01</td>
<td>5.17</td>
<td>4.91</td>
<td>5.11</td>
<td>2.29</td>
</tr>
<tr>
<td>RLPrompt</td>
<td>6.76</td>
<td>4.09</td>
<td>4.73</td>
<td>5.27</td>
<td>5.74</td>
<td>4.99</td>
<td>5.16</td>
<td>4.87</td>
<td>5.10</td>
<td>2.28</td>
</tr>
</tbody>
</table>

Selecting most sensitive instruction selection methods:
- Task-agnostic instructions dominate few-shot settings.
- Automatic instruction selection methods outperformed by simple baselines!

**Takeaways**

Existing automatic instruction selection methods
- do not generalize well to more models and tasks.
- may require extensive hyperparameter tuning.
- can be computationally 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 (e.g., 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**