**SKILL-MIX Evaluation**

- **Set of N language skills**
  - (Example: metaphor)
  - Not rare in training set (e.g., in Wikipedia)
- **List of T topics**
  - (Example: sewing)
  - Low, non-negligible probability
  - Pick random subset of k skills and 1 topic
- **Ask model (Student) to output a short piece of text**
  - in the context of the topic and illustrate all k skills.
- **Second attempt is usually better**
- **Motivations:** [2] gives a theory for emergence on complex skills
  - Predicts that scaling model by 10x doubles k for which model is successful

---

**SKILL-MIX Pipeline**

- **Generate**
  - Given k, skill, and topic combinations
- **Part 1: Generation**
  - Generated Text (Round 1)
  - Generated Text (Round 2)
- **Part 2: Grading**
  - Get Aggregated Grade
  - Get Aggregated Grade
- **Report average over all scores**

---

**SKILL-MIX: A Flexible and Expandable Family of Evaluations for AI Models**

Dingli Yu† • Simran Kaur† • Arushi Gupta† • Jonath Brown-Cohen† • Anirudh Goyal‡ • Sanjeev Arora‡

† Princeton Language and Intelligence (PLI), Princeton University  ‡Google DeepMind

**Issues in evaluating LLMs**

- Data contamination
- Model originality
- Cramping for leaderboards
- Debate: are LLMs “stochastic parrots”? [1]

**Desiderata for next-generation evaluations**

- Relevant to general-purpose intelligence + language understanding
- Easy to design and evaluate
- Resistant to data contamination
- Capable of revealing the novelty (go beyond “stochastic parrots” behavior)
- Resistant to “cramping for leaderboard”
- Easy to adjust the difficulty level, and applicable to all models

---

**Metrics**

- Ratio of Full Marks:
  - 1 f all k + 3 points are earned, and 0 otherwise
- Ratio of All Skills: 1 f k points are awarded for k skills and ≥ 2 points are awarded for remaining criteria, and 0 otherwise
- **Skill Fraction:** points awarded for k skills k
  - if all 3 points are awarded for remaining criteria, and 0 otherwise

---

**SKILL-MIX as grader**

<table>
<thead>
<tr>
<th>Model (generation)</th>
<th>k</th>
<th>k−1</th>
<th>k-2</th>
<th>k-3</th>
<th>k-4</th>
<th>k-5</th>
<th>k-6</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLaMA-2-7B-Chat</td>
<td>0.50</td>
<td>0.65</td>
<td>0.76</td>
<td>0.79</td>
<td>0.80</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>LLaMA-2-13B-Chat</td>
<td>0.50</td>
<td>0.65</td>
<td>0.76</td>
<td>0.79</td>
<td>0.80</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>LLaMA-2-70B-Chat</td>
<td>0.50</td>
<td>0.65</td>
<td>0.76</td>
<td>0.79</td>
<td>0.80</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>GPT-4</td>
<td>0.50</td>
<td>0.65</td>
<td>0.76</td>
<td>0.79</td>
<td>0.80</td>
<td>0.84</td>
<td></td>
</tr>
</tbody>
</table>

**Llama-2-70A as grader**

<table>
<thead>
<tr>
<th>Model (generation)</th>
<th>k</th>
<th>k−1</th>
<th>k-2</th>
<th>k-3</th>
<th>k-4</th>
<th>k-5</th>
<th>k-6</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLaMA-2-7B-Chat</td>
<td>0.50</td>
<td>0.65</td>
<td>0.76</td>
<td>0.79</td>
<td>0.80</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>LLaMA-2-13B-Chat</td>
<td>0.50</td>
<td>0.65</td>
<td>0.76</td>
<td>0.79</td>
<td>0.80</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>LLaMA-2-70B-Chat</td>
<td>0.50</td>
<td>0.65</td>
<td>0.76</td>
<td>0.79</td>
<td>0.80</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>GPT-4</td>
<td>0.50</td>
<td>0.65</td>
<td>0.76</td>
<td>0.79</td>
<td>0.80</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>Mistral-7B-Instruct</td>
<td>0.50</td>
<td>0.65</td>
<td>0.76</td>
<td>0.79</td>
<td>0.80</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>Qwen-14B-Chat</td>
<td>0.50</td>
<td>0.65</td>
<td>0.76</td>
<td>0.79</td>
<td>0.80</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>Xwin-LM-70B-V0.1</td>
<td>0.50</td>
<td>0.65</td>
<td>0.76</td>
<td>0.79</td>
<td>0.80</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>Falcon-10B-Chat</td>
<td>0.50</td>
<td>0.65</td>
<td>0.76</td>
<td>0.79</td>
<td>0.80</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>Tigereye-10B-Chat</td>
<td>0.50</td>
<td>0.65</td>
<td>0.76</td>
<td>0.79</td>
<td>0.80</td>
<td>0.84</td>
<td></td>
</tr>
</tbody>
</table>

**Beyond Stochastic Parrot Behavior**

- Let p_r be the average frequency of skills, p_t be the average frequency of topics, L be the length of training corpus
- The total occurrence of all possible combination in training is

\[\max_{\pi_1, \ldots, \pi_L} \prod_{i=1}^{L} p_{r_i} \prod_{j=1}^{L} \sum_{\pi_j} p_{t_{\pi_j}} \leq \prod_{i=1}^{L} p_{r_i} \sum_{\pi_j} p_{t_{\pi_j}} \]

- which is bounded by \(p_r p_t^{L-1} = \frac{1}{T} \) for k = 5
- **GPT-4 can output ~12% of \(\left(\frac{2}{3}\right)^k\) for k = 5, ~1% for k = 6

**Conclusion**

- Evaluation is hard because language has a long-tailed distribution.
- SKILL-MIX accesses this long-tail by random combinations of k skills.
- **SKILL-MIX tests a model’s ability to describe/imagine situations that do not exactly match anything seen during training.**
- **Performance of models on SKILL-MIX generally accord with popular perceptions of their quality.**
- **GPT-4 surpasses “stochastic parrot” behavior** [1]

---