Beyond Testers’ Biases: Guiding Model Testing with Knowledge Bases using LLMs

Chenyang Yang, Rishabh Rustogi, Rachel Brower-Sinning, Grace Lewis, Christian Kaestner, Tongshuang (Sherry) Wu
Coarse-grained Model Evaluation is not Enough

Model evaluations use a single score to compare and rank different ML models. Coarse-grained evaluations can not provide insights for models’ strengths and weaknesses, which are useful for model analysis, debugging, and selection.
# Beyond Accuracy: Behavioral Model Testing

<table>
<thead>
<tr>
<th>Capabilities</th>
<th>Descriptions</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocab/POS</td>
<td>important words or word types for the task.</td>
<td>template('This is a {adj:mask} movie.')</td>
</tr>
<tr>
<td>Named entities</td>
<td>appropriately understanding named entities.</td>
<td>perturb('[John] doesn’t like the movie', change_name)</td>
</tr>
<tr>
<td>Negation</td>
<td>understand the negation words.</td>
<td>template('The food is not {adj:mask}.')</td>
</tr>
<tr>
<td>Taxonomy</td>
<td>synonyms, antonyms, etc.</td>
<td>perturb('How can I become more {optimistic}?', antonym)</td>
</tr>
<tr>
<td>Robustness</td>
<td>to typos, irrelevant changes, etc.</td>
<td>perturb('@SouthwestAir no {thanks}', replace_char)</td>
</tr>
<tr>
<td>Coreference</td>
<td>resolve ambiguous pronouns, etc.</td>
<td></td>
</tr>
<tr>
<td>Fairness</td>
<td>not biasing towards certain gender/race groups.</td>
<td></td>
</tr>
<tr>
<td>Semantic Role Labeling</td>
<td>understanding roles such as agent, object, etc.</td>
<td></td>
</tr>
<tr>
<td>Logic</td>
<td>handle symmetry, consistency, and conjunctions.</td>
<td></td>
</tr>
<tr>
<td>Temporal</td>
<td>understand order of events.</td>
<td>template('I used to hate the {noun:mask}, but now I like it')</td>
</tr>
</tbody>
</table>

CheckList: Applying the **principles for software testing** to model testing. Create test cases for **concrete model behavior**.

[ACL'20] Beyond Accuracy: Behavioral Testing of NLP Models with CheckList
Beyond Accuracy: Behavioral Model Testing

CheckList: Applying the **principles for software testing** to model testing.

**What to test:** Pre-defined capabilities

**How to test:** Templates, perturbations

**Test oracle:** Specified outputs, metamorphic relations

AdaTest: Use **LLMs** to **suggest** test cases for user-defined capabilities.

**What to test:** User-defined

**How to test:** LLM suggestions

**Test oracle:** Specified outputs

Different work varies on these three dimensions
Model Testing is Ad-hoc and Biased

CheckList: Applying the principles for software testing to model testing.

**What to test:** Pre-defined capabilities

**How to test:** Templates, perturbations

**What to test:** User-defined

**How to test:** LLM suggestions

AdaTest: Use LLMs to suggest test cases for user-defined capabilities.

Existing model testing methods focus on how to test, exploring different test generation methods. But how do testers know what to test?
Model Testing is Ad-hoc and Biased

Users tend to explore **locally**
overfit to their intuition, domain knowledge, confirmation bias.

Expect: **Comprehensive testing**
More systematically cover the space beyond individual biases
Model Testing is Ad-hoc and Biased

How to move from **local exploitation** towards **global exploration**?

Users tend to explore **locally** overfit to their intuition, domain knowledge, confirmation bias.

Expect: **Comprehensive testing**
More systematically cover the space beyond individual biases.
SE to NLP: Requirements Engineering

V-Model: software design & verification, grounded on requirements
SE to NLP: Requirements Engineering

**Weaver:** Help users test models for their specific tasks comprehensively, by helping them elicit relevant requirements with LLM-generated knowledge base.

**V-Model:** software design & verification, grounded on requirements
**Weaver Workflow**

**Seed concept:** Online toxicity

**Query LLMs for concepts** (ConceptNet relations)
- **MannerOf:** List some ways to do online toxicity: Harassment, Cyberbullying...
- **TypesOf:** List some types of online toxicity: Racism, Misogyny...

**Resulting LLM-generated Knowledge base**

Intuition: LLMs have parametric knowledge for various domains, tasks, and topics.

Traditional knowledge base relation helps extract the knowledge comprehensively.
**Seed concept:** Online toxicity

**Query LLMs for concepts** (ConceptNet relations)

- **MannerOf:** List some ways to do online toxicity: Harassment, Cyberbullying...
- **TypesOf:** List some types of online toxicity: Racism, Misogyny...

**Resulting LLM-generated Knowledge base**

- Harassment
- Cyberbullying
- Deceitful statement
- Conspiracy theory
- Spreading misinformation
- Online toxicity
- Hate speech
- Religious discrimination
- Misogyny
- Racism
- Disabled people
- Immigrant
- Religions

**Example Panel**

- Concept: Misogyny
  
  **Path:** Misogyny is a type of online toxicity.

  **Suggestions**  **Add Examples**

  **Ask for suggestions...**
Weaver recommends important concepts

Comparing **Weaver concepts** vs. **gold concepts** identified from existing dataset analysis and user studies

<table>
<thead>
<tr>
<th>Task</th>
<th>Recall</th>
<th>Precision</th>
<th># Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hateful meme detection</td>
<td>93.1%</td>
<td>88.0%</td>
<td>101</td>
</tr>
<tr>
<td>Pedestrian detection</td>
<td>91.8%</td>
<td>74.0%</td>
<td>146</td>
</tr>
<tr>
<td>Stance detection for feminism</td>
<td>86.9%</td>
<td>84.0%</td>
<td>145</td>
</tr>
<tr>
<td>Stance det. for climate change</td>
<td>91.4%</td>
<td>76.0%</td>
<td>185</td>
</tr>
<tr>
<td>Average</td>
<td>90.6%</td>
<td>80.5%</td>
<td>144</td>
</tr>
</tbody>
</table>

**Weaver covers most of the important concepts** even when we only grow the knowledge base to the second layer!
Weaver supports systematic bug finding

We conducted a within-subject controlled experiment (N=20) to see whether Weaver helps users...

- Explore more concepts?
- Explore more diverse concepts?
- Mitigate their biases?
Weaver supports systematic bug finding vs. Manually adding concepts while exploring model errors (on LLM-generated inputs), Weaver...

Helps identify 57.5% more concepts in the same amount of time
Weaver supports systematic bug finding

vs. Manually adding concepts while exploring model errors (on LLM-generated inputs), Weaver…

These concepts are more diverse
(covered 47.7% more clusters)

Helps identify **57.5% more concepts** in the same amount of time
Weaver supports systematic bug finding vs. Manually adding concepts while exploring model errors (on LLM-generated inputs), Weaver...

Enables users to continuously discover distinct concepts (vs. control: focus on refining existing concepts)

Helps identify 57.5% more concepts in the same amount of time
Weaver supports systematic bug finding vs. Manually adding concepts while exploring model errors (on LLM-generated inputs), Weaver…

Enables users to explore concepts beyond their biases

- **[Violence]**: Verbal abuse of women is not as bad as physical violence.
  - Favor: against
- **[Sexual Harassment]**: Sexual harassment in the workplace against women should not be tolerated. Sexual harassment against men is not a big deal. #MeToo
  - Against: favor
- **[STEM]**: Women are underrepresented in STEM due to a lack of access to education and resources.
  - Favor: none
- **[Admission]**: It’s not important to ensure that admissions processes are free of gender bias and discrimination.
  - Against: favor
We conducted two case studies to see whether Weaver is useful in real-world settings.

Weaver helps practitioners test (and iterate) prompts

C1: Prompt LLMs to summarize transcripts

Weaver helps practitioners find new bugs

“Summaries are chronological even when reordering is desired”

C2: Prompt LLMs to explain code

“Specific challenges that novice programmers might have in comprehending [domain] code”

Weaver can help early-stage development
We conducted two case studies to see whether Weaver is useful in real-world settings.

**Weaver helps practitioners test (and iterate) prompts**

C1: Prompt LLMs to summarize transcripts

C2: Prompt LLMs to explain code

Looking for more users!
Weaver helps users & practitioners...
identify more concepts and more diverse concepts
find bugs beyond their biases
test (and iterate) prompts in domain applications