EvalAgent: Discovering Implicit Evaluation Criteria from the Web

TAUR LAB @UT Austin

Manya Wadhwa, Zayne Sprague, Chaitanya Malaviya, Philippe Laban, Junyi Jessy Li, Greg Durrett

manyawadhwa@nyu.edu (work done at UT Austin)

Paper

Given a prompt, EvalAgent finds important yet non-obvious evaluation criteria.

Our goal: identifying evaluation criteria

🙂 *Write a New Yorker style fiction piece given the plot below: […]*

We could evaluate responses to this prompt with these **criteria:**

1. Response should be a New Yorker style fiction piece
2. Response should have details about the plot and characters

Instruction decomposition finds these **obvious criteria**. LLMs can generate other criteria:

1. Response should draw an eerie connection between the woman and a discarded chair (plot)
2. The response should develop characters & themes according to literary fiction standards

These are often vague and not actionable. **Hard to evaluate and fix responses.**

This work: we automatically identify criteria that are:

- **Specific**: criterion is a precise dimension of quality
- **Non-obvious**: criterion is implicit, unspoken principle
- **Actionable**: criterion ensures tangible improvements

EvalAgent: discovering specific, non-obvious criteria from instructional web documents

Step 1: Generate queries to retrieve relevant how-to docs

`how to write good fiction`

`Query 2`

`Query n`

Step 2: Retrieve *instructional* web documents, filter and summarize them into query-specific criteria

https://writers.com/how-to-start-writing-fiction

*[….] Consider starting the story with **what makes your world live**: a pulsing city, the whispered susurrus of orchards, hills that roil with unsolved mysteries, etc. Tell us where the conflict is happening, and the story will follow.[…]*

https://www.newyorker.com/humor/daily-shouts/eight-rules-for-writing-fiction

*[….] **Give your characters motivations.** Ask yourself in each scene, "What does this character want?"[…]*

Step 3: Combine all query specific criteria, filter and rewrite it to be task-aligned

The response should:
1. Have settings with sensory details
2. Tell a story with narrative voice and tone
3. Have characters with goals
4. Exhibit character actions with consequences
5. Have characters reveal traits through dialogue
6. Have varied pacing to reflect character states

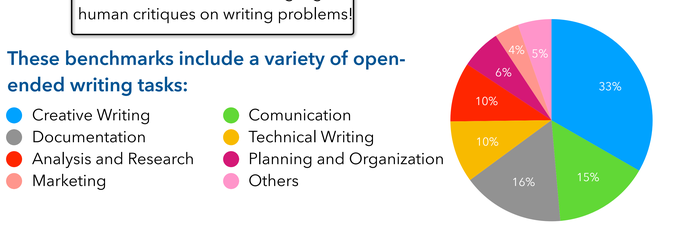
Criteria by EvalAgent!

Benchmark Datasets (+ a new one!)

Datasets	Criteria Source	Expert oversight
BigGenBench (Kim et al, 2024)	Human	✘
Art or Artifice (Chakrabarty et al, 2024)	Human	✓
Dolomites (Malaviya et al, 2025)	Human	✓
InfoBench (Qin et al, 2024)	Human	✓
MT-Bench (Zheng et al, 2023)	LLM	✘
WildBench (Lin et al, 2024)	LLM	✘
WritingBench (Wu et al, 2025)	LLM	✘
Ask-then-critique!	Human	✓

New dataset with natural language human critiques on writing problems!

These benchmarks include a variety of open-ended writing tasks:



For more results on how we can integrate EvalAgent with existing methods and check alignment with human criteria, check out the paper!

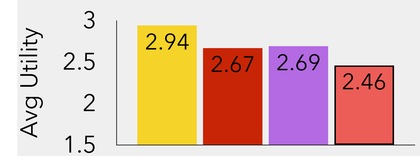
Evaluation Task: generate useful evaluation criteria for open-ended tasks (writing, etc.)

Baseline methods: (1) Instruction decomposition (ID): break an instruction down into criteria. (2) Prompting an LLM to generate criteria.

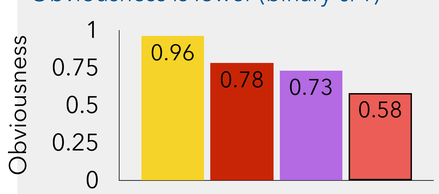
Results: ID LLM Human EvalAgent

Human ratings of EvalAgent criteria

Utility is high (1-3 scale); the criteria are useful!

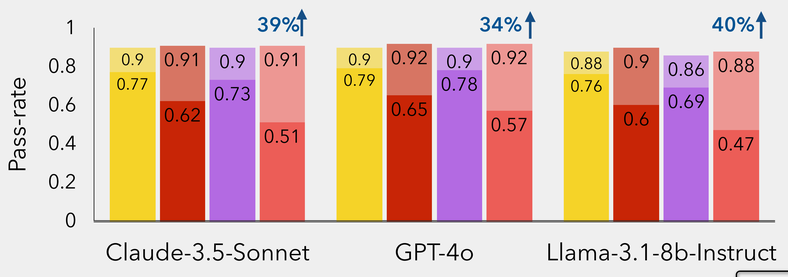


Obviousness is lower (binary 0/1)



Pass rate of responses when judged against different criteria

LLM responses score lower on EvalAgent criteria compared to others (non-obvious, not saturated)
Refining with EvalAgent criteria leads to large gains (actionable)



LLMs fail to satisfy implicit and non-obvious criteria