The BIGGEN BENCH: A Principled Benchmark for Fine-grained Evaluation of Language Models with Language Models

Seungone Kim[1*] Juyoung Suk[2] Ji Yong Cho[3,4] Shayne Longpre[5]
Chaeeun Kim[2] Dongkeun Yoon[2] Guijin Son[6] Yejin Cho[2] Sheikh Shafayat[2]
Jinheon Baek[2] Sue Hyun Park[2] Hyeonbin Hwang[2] Jinkyung Jo[2] Hyowon Cho[2]
Haebin Shin[2] Seongyun Lee[2] Hanseok Oh[2] Noah Lee[2] Namgyu Ho[2]
Se June Joo[2] Miyoung Ko[2] Yoonjoo Lee[2] Hyungjoo Chae[6]
Jamin Shin[2] Joel Jang[7] Seonghyeon Ye[2] Bill Yuchen Lin[7]
Sean Welleck[1] Graham Neubig[1] Moontae Lee[3,8] Kyungjae Lee[3] Minjoon Seo[2]

[1] Carnegie Mellon University [2] KAIST [3] LG AI Research [4] Cornell University
[5] MIT [6] Yonsei University [7] University of Washington [8] University of Illinois Chicago

NAACL 2025 Best Paper Award

HUMANE Lab 박현빈

25.07.09

Background

- Most generation benchmarks currently assess LMs using abstract evaluation criteria – like helpfulness and harmlessness
- Additionally, these benchmarks tend to focus disproportionately on specific capabilities such as instruction following

Evaluation Criteria

[Input Prompt]

Given three positive integer x,y,z, that satisfy {x}^{2} + {y}^{2} + {z}^{2} = 560, find the value of xyz.
You are not allowed to use your code functionality.

Coarse-grained Evaluation Criteria

Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses.

Domain-Specific Evaluation Criteria

Score 1 The logic of the model's response is completely incoherent.
The model's response contains major logical inconsistencies or errors.
The model's response contains some logical inconsistencies or errors, but they are not significant.
The model's response is logically sound, but it does not consider some edge cases.
Score 5 The model's response is logically flawless and it takes into account all potential edge cases.

Instance-Specific Evaluation Criteria

Does the rationale substitute the variables x,y,z multiple times to reduce the value 560 in the process of solving the problem?

Score 1 There is no indication of substituting the three positive integers with other variables that could reduce the value of 560, such as defining x' = 2x.

Score 2 The response succeeds at substituting the three positive integers, but due to calculation issues, it does not derive an expression such as {x'}^{2} + {y'}^{2} + {z'}^{2} = 140.

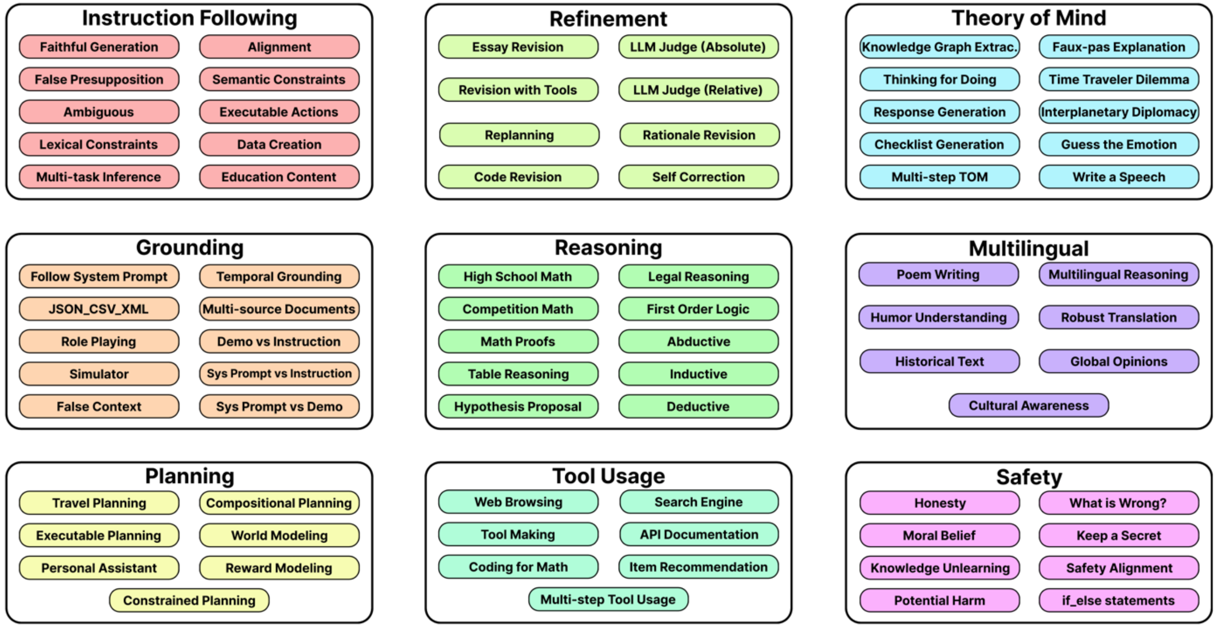
Score 3 After acquiring an expression similar to {x'}^{2} + {y'}^{2} + {z'}^{2} = 140, the response fails to apply the same logic once more and acquire an expression such as {x''}^{2} + {y''}^{2} + {z''}^{2} = 35.

Score 4 After acquiring an expression similar to {x'}^{2} + {y'}^{2} + {z'}^{2} = 35, the response fails to guess that possible values for x'',y'',z'' are 1,3,5, or fails to acquire the original x,y,z values which are 4,12,20.

Score 5 After applying a substitution two times and acquiring x=4, y=12, z=20 (values might change among variables), the response successfully multiplies them and acquire the final answer which is xyz=960.

The BigGen Bench

- Evaluates 9 capabilities of LMs across 77 tasks and 765 instances
- LLM-as-a-judge
- All instances were crafted through a human-in-the-loop approach

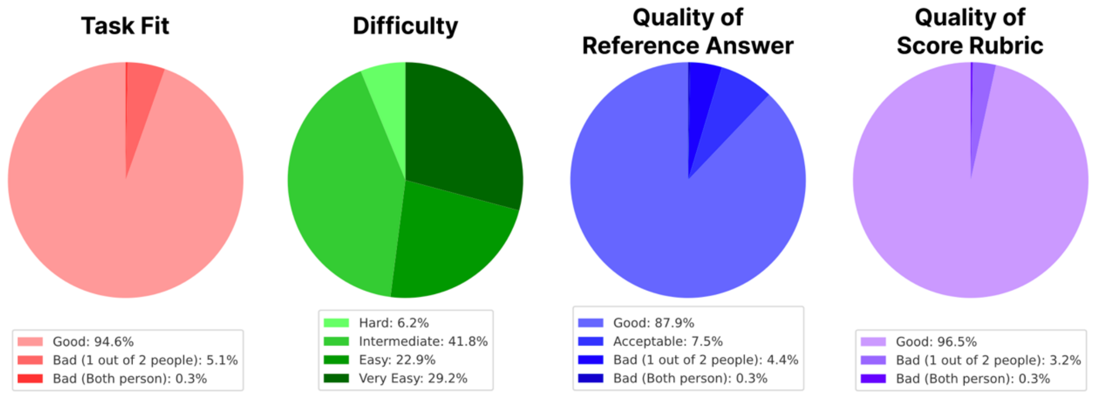


Construction Process

- Each instance includes a system message, an input, a reference answer, and a scoring rubric
- Step1: Hand-crafting instances → 385
- Step2: Augmenting new instances with human demonstrations → 770

  Creating new instances by prompting GPT-4 with Step 1 instances

- Step3: Cross validation → 765



- Step4: Gathering human judgements

  gathering human judgements to verify the reliability of evaluation results from judge LMs

Evaluation Protocol

- Using zero-shot prompting if the Response LM is post-trained
- Using 3-shot prompting if the Response LM is pre-trained
- Evaluator LM takes in a single response from the response LM

```
###Task Description:
An instruction (might include an
    Input inside it), a response
    to evaluate, a reference
    answer that gets a score of 5,
     and a score rubric
    representing a evaluation
    criteria are given.
1. Write a detailed feedback that
    assess the quality of the
    response strictly based on the
     given score rubric, not
    evaluating in general.
2. After writing a feedback, write
     a score that is an integer
    between 1 and 5. You should
    refer to the score rubric.
3. The output format should look
    as follows: "Feedback: (write
    a feedback for criteria) [
    RESULT] (an integer number
    between 1 and 5)"
4. Please do not generate any
    other opening, closing, and
    explanations.
```

```
###The instruction to evaluate:
{orig_instruction}

###Response to evaluate:
{orig_response}

###Reference Answer (Score 5):
{orig_reference_answer}

###Score Rubrics:
{score_rubric}

###Feedback:
```

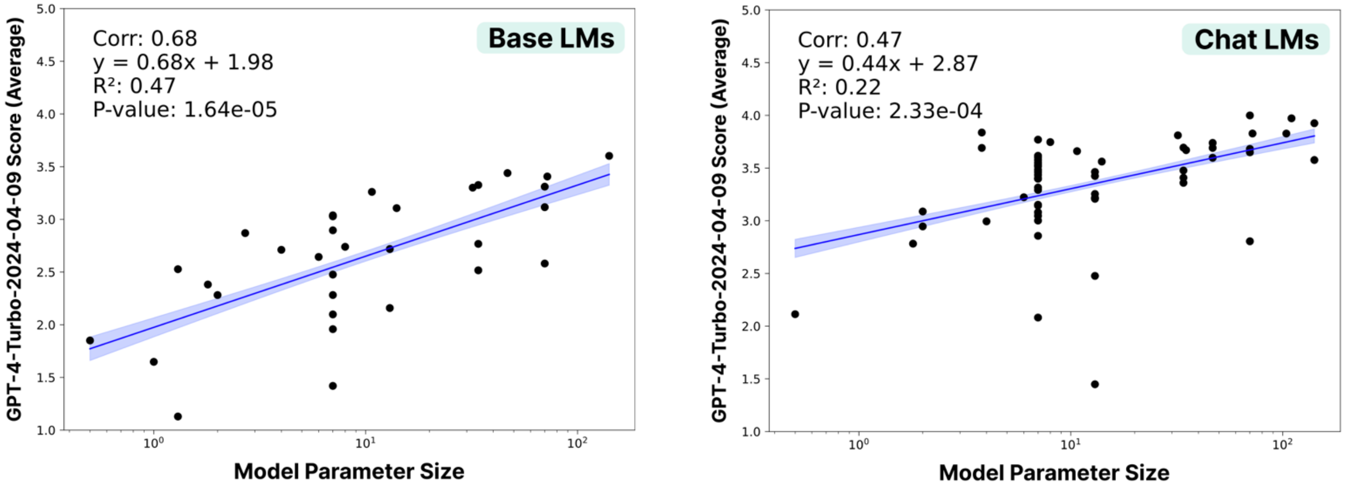
Main Results and Analyses

- Evaluate 103 LMs using 5 different judge LMs
  - GPT-4-1106
  - GPT-4-2024-04-09
  - Prometheus-2-8x7B
  - Prometheus-2-8x7B-BGB
  - Claude-3-Opus

model_name	grounding	instruction_following	planning	reasoning	refinement	safety	theory_of_mind	tool_usage	multilingual
phi-1	1.100	1.000	1.000	1.000	1.303	1.391	1.010	1.012	nan
phi-1_5	2.425	2.770	2.314	2.130	2.329	2.870	2.700	1.300	nan
phi-2	3.050	2.860	2.600	2.700	2.789	3.406	3.000	1.675	nan
Qwen1.5-0.5B	1.850	2.060	1.471	1.500	1.934	2.029	1.750	1.150	nan
Qwen1.5-1.8B	2.425	2.790	2.214	1.830	2.408	2.420	2.360	1.413	nan
Qwen1.5-4B	2.850	2.820	2.557	2.300	2.447	3.130	2.610	1.688	nan
gemma-2b	2.163	2.610	2.129	1.990	1.934	2.420	2.240	1.350	nan
OLMo-1B	1.675	1.700	1.343	1.330	1.737	2.072	1.440	1.087	nan
Qwen1.5-0.5B-Chat	2.075	2.360	1.957	1.680	1.776	2.594	2.260	1.250	1.116
Qwen1.5-1.8B-Chat	2.750	3.090	2.629	2.280	2.553	2.696	3.030	1.688	1.314
Qwen1.5-4B-Chat	2.862	2.990	2.914	2.690	2.579	3.362	2.890	2.050	1.400
Phi-3-mini-4k-instruct	3.675	3.820	3.486	3.590	3.763	4.101	3.780	3.112	1.743
Phi-3-mini-128k-instruct	3.500	3.660	3.500	3.610	3.539	3.986	3.660	2.700	1.743
gemma-2b-it	2.825	3.120	3.000	2.390	2.724	3.928	3.160	1.812	1.514
gemma-1.1-2b-it	2.812	3.210	3.000	2.490	2.947	3.884	3.150	1.675	1.386
gemma-7b	1.288	1.530	1.171	1.280	1.474	2.029	1.170	1.025	nan
Mistral-7B-v0.1	3.150	3.220	3.029	2.750	2.566	3.290	2.970	2.038	nan

Main Results and Analyses

- Performance of base LMs increases smoothly with scaling model parameter size
- Performance of chat LMs is not only attributed to model size scaling



Main Results and Analyses

- The performance gap closes between larger base and chat LMs, remains in smaller models
- Identifying performance gap between open-source and proprietary LMs

Capability	**Coefficient**
Average	-0.08^{***}
Refinement	-0.05^{***}
Reasoning	-0.07^{***}
Grounding	-0.07^{***}
Planning	-0.07^{***}
Tool Usage	-0.08^{***}
Safety	-0.09^{***}
Instruction Following	-0.09^{***}
Theory of Mind	-0.14^{***}

Capability	**Hedges's *g***
Average	0.51
Safety	0.36
Instruction Following	0.38
Refinement	0.46
Grounding	0.49
Tool Usage	0.58
Planning	0.58
Theory of Mind	0.59
Reasoning	0.65
Multilingual	0.84

Can We Rely on Judge Models?

- Judge LMs can mimic human judgement

Evalautor LM	Inst. Follow.	Ground.	Reason.	Plan.	Refine.	Multi.	Safety	ToM	Tool.	Average
Prometheus-2 8x7B	0.413	0.526	0.517	0.607	0.421	0.459	0.516	0.371	0.412	0.471
+ *Self-Consistency* (N=3)	0.432	0.583	0.549	0.590	0.455	0.502	0.571	0.371	0.469	0.502
+ *Self-Consistency* (N=5)	0.465	0.577	0.539	0.593	0.436	0.484	0.593	0.392	0.452	0.503
Prometheus-2-BGB 8x7B	0.620	0.661	0.626	0.642	0.516	0.554	0.691	0.441	0.441	0.577
+ *Self-Consistency* (N=3)	0.643	0.699	0.665	0.701	0.585	0.540	0.678	0.501	0.455	0.607
+ *Self-Consistency* (N=5)	0.619	0.689	0.659	0.716	0.577	0.545	0.672	0.533	0.455	0.607
Claude-3-Opus	0.624	0.694	0.588	0.634	0.561	0.554	0.634	0.463	0.446	0.578
GPT-4-1106	0.641	0.683	0.643	0.678	0.578	0.583	0.653	0.420	0.496	0.597
GPT-4-Turbo-2024-04-09	0.647	0.718	0.695	0.678	0.578	0.574	0.692	0.478	0.551	0.623
Majority Voting	0.646	0.715	0.674	0.708	0.575	0.611	0.687	0.497	0.529	0.627

pearson correlation between judge LMs and human evaluators

Can We Rely on Judge Models?

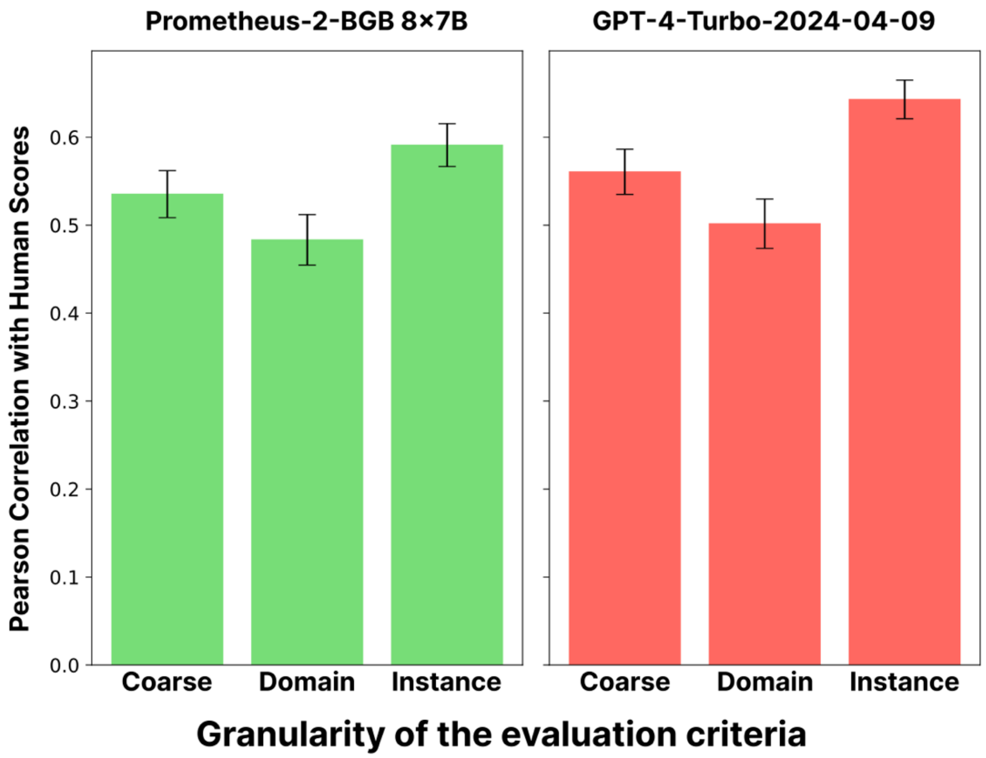
- As Prometheus-2-BGB trained on BigGen Bench, it is questionable whether its evaluation performance decreases when assessing other benchmarks

Evaluator LM	Vicuna Bench		MT Bench		FLASK			Feedback Bench
	GPT-4-1106	Claude-3-Opus	GPT-4-1106	Claude-3-Opus	GPT-4-1106	Claude-3-Opus	Humans	GPT-4-0613
Mistral-Instruct-7B	0.486	0.561	0.284	0.396	0.448	0.437	0.377	0.586
Mixtral-Instruct-8x7B	0.566	0.579	0.551	0.539	0.483	0.495	0.420	0.673
Prometheus-2-7B	0.642	0.610	0.543	0.554	0.645	0.578	0.544	0.878
Prometheus-2-8x7B	<u>0.685</u>	**0.635**	<u>0.665</u>	<u>0.614</u>	<u>0.659</u>	<u>0.626</u>	<u>0.555</u>	**0.898**
Prometheus-2-BGB-8x7B (Ours)	**0.777**	<u>0.618</u>	**0.773**	**0.619**	**0.764**	**0.635**	**0.649**	<u>0.890</u>
GPT-3.5-Turbo-0613	0.335	0.349	0.183	0.194	0.437	0.396	0.450	0.594
GPT-4-1106	/	0.694	/	0.717	/	0.736	0.679	0.753
Claude-3-Opus	0.694	/	0.717	/	0.736	/	0.573	0.788

pearson correlation between evaluator LMs

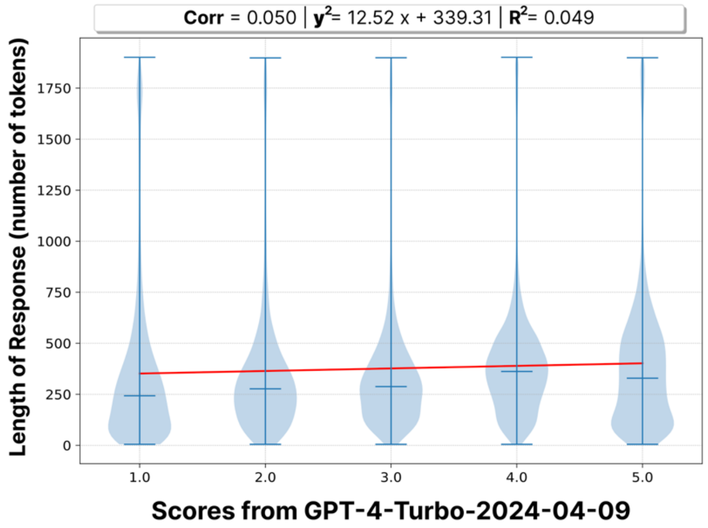
Are Fine-Grained Criteria Crucial?

- Instance-specific criteria enable accurate judgements



Analysis of Verbosity Bias

- Analyzing whether evaluator LMs prefer longer responses



- Correlation coefficient of 0.05 and an R^2 value of 0.049 indicate a very weak linear relationship
- Due to a detailed scoring rubric and direct assessment, the impact of longer responses was slight.

My Review

- Instance-specific rubric
- High quality human-in-the-loop design
- Creating a powerful open-source judge model
- Capability Gap Analysis
- High potential for future adoption