A Task-Level Case Study

This section illustrates how a model's performance may vary across different tasks associated with the same new term. We analyzed the performance of Llama-3-Instruct-70B on the new term "wokely," defined as an adjective meaning "Of little worth; poor, mean, paltry." The model's performance varied across three tasks under the zero-shot Base setting:

Task	**Question**	**Response**
COMA	The book's cover was described as wokely by several reviewers. I am hesitating among these options. Help me choose the more likely effect: A. it struggled to attract attention on the bookstore displays despite a compelling narrative inside. B. many readers were enticed to buy it, strengthening its presence on the bestseller list. C. readers were intrigued and the book's sales experienced an unexpected surge worldwide. D. the publisher decided to release a limited edition with a special hardback velvet cover.	A (✓)
COST	The goods at the flea market appeared distinctly _, making it hard to find a satisfying purchase. In the previous sentence, does _ refer to A. Spokely, B. Cokely, C. Wokely, or D. Worthy?	D (X)
CSJ	His contributions to the project were considered wokely, barely making any impact. Is this example in line with commonsense and grammatically correct?	Incorrect (X)

Table 2: Performance of Llama-3-Instruct-70B on Different Tasks Involving the New Term "wokely"

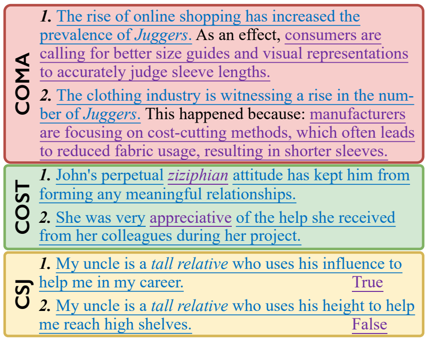
As observed, the model only answered correctly in the COMA task but failed in the other two tasks. In the COMA task, the model successfully inferred that "wokely" carries a negative connotation, allowing it to correctly choose choice A. This demonstrates its ability to *comprehend* the new term within a *helpful* context. However, in the COST task, where the model needed to *utilize* the new term and *distinguish* it from similar choices, it struggled. Although the phrase "hard to find a satisfying purchase" suggested the need for a negative term, the model incorrectly chose "Worthy," which is grammatically correct but semantically incorrect. In the CSJ task, the model was required to *process* and *interpret* the new term in the *absence* of helpful *context*. The context matched the definition of "wokely" perfectly, yet the model erroneously judged the response as incorrect because it was a judgment-based evaluation.

These results provide a comprehensive evaluation of the model's understanding of the term "wokely." They reveal that while the model can recognize that it is a negative term when the context is clear, it struggles to grasp the detailed meaning of the term and how to accurately use it in different contexts.

B Benchmark Generation Cases and Prompts

Benchmark generation cases. For clarity, we provide cases to illustrate how to extract questions and correct choices from sentences, as shown in Figure 7. In these examples, the two cases from COMA correspond to the inclusion of fixed phrases "As an effect" and "This happened because", respectively. The two cases from COST represent the new term and its related term as the correct answers, respectively. The two cases from CSJ correspond to questions with answers being True and those with modified answers being False, respectively.

Furthermore, we provide an example of the COMA task construction process, as shown in Figure 8. Ultimately, we filter out choices A and E, resulting in the final clean question being the current question, along with a multiple choice question that contains only choices B, C, D and F.



1. The rise of online shopping has increased the prevalence of *Juggers*. As an effect, consumers are calling for better size guides and visual representations to accurately judge sleeve lengths.

2. The clothing industry is witnessing a rise in the number of *Juggers*. This happened because: manufacturers are focusing on cost-cutting methods, which often leads to reduced fabric usage, resulting in shorter sleeves.

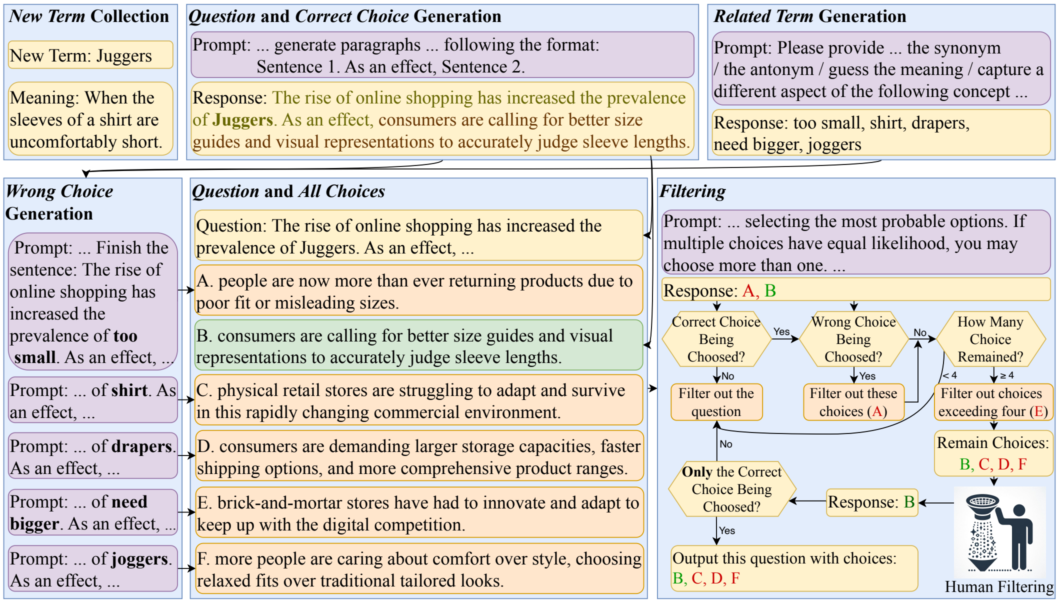
1. John's perpetual *ziziphian* attitude has kept him from forming any meaningful relationships.

2. She was very appreciative of the help she received from her colleagues during her project.

1. My uncle is a *tall relative* who uses his influence to help me in my career. True

2. My uncle is a *tall relative* who uses his height to help me reach high shelves. False

Figure 7: Examples of question and correct choice generation. We first generate sentence, then separate it to obtain the question and correct choice.


***New Term* Collection**
Question and Correct Choice Generation
***Related Term* Generation**
***Wrong Choice* Generation**
Question* and *All Choices
Filtering
Correct Choice Being Choosed?
Yes
No
Wrong Choice Being Choosed?
Yes
No
How Many Choice Remained?
< 4
≥ 4
Filter out the question
Filter out these choices (A)
Filter out choices exceeding four (E)
Remain Choices: B, C, D, F
No
Only the Correct Choice Being Choosed?
Yes
Response: B
Human Filtering

New Term: Juggers

Meaning: When the sleeves of a shirt are uncomfortably short.

Prompt: ... generate paragraphs ... following the format: Sentence 1. As an effect, Sentence 2.

Response: The rise of online shopping has increased the prevalence of **Juggers**. As an effect, consumers are calling for better size guides and visual representations to accurately judge sleeve lengths.

Prompt: Please provide ... the synonym / the antonym / guess the meaning / capture a different aspect of the following concept ...

Response: too small, shirt, drapers, need bigger, joggers

Prompt: ... Finish the sentence: The rise of online shopping has increased the prevalence of **too small**. As an effect, ...

Prompt: ... of **shirt**. As an effect, ...

Prompt: ... of **drapers**. As an effect, ...

Prompt: ... of **need bigger**. As an effect, ...

Prompt: ... of **joggers**. As an effect, ...

Question: The rise of online shopping has increased the prevalence of Juggers. As an effect, ...

A. people are now more than ever returning products due to poor fit or misleading sizes.

B. consumers are calling for better size guides and visual representations to accurately judge sleeve lengths.

C. physical retail stores are struggling to adapt and survive in this rapidly changing commercial environment.

D. consumers are demanding larger storage capacities, faster shipping options, and more comprehensive product ranges.

E. brick-and-mortar stores have had to innovate and adapt to keep up with the digital competition.

F. more people are caring about comfort over style, choosing relaxed fits over traditional tailored looks.

Prompt: ... selecting the most probable options. If multiple choices have equal likelihood, you may choose more than one. ...

Response: A, B

Output this question with choices: B, C, D, F

Figure 8: An example of the COMA task construction process. The input is the collected new term and its meaning, and the output is the question with choices B, C, D, and F, where B is the correct choice.

Benchmark generation prompts. Further, we introduce the prompt we used in benchmark construction and LLM evaluation. We use "[·]" to express variables depending on the input. The notation "[W]" represents the new term and "[M]" represents the meaning of the term. We use "[Ti]" to represent the i-th related term of the new term, "[Ci]" to represent the i-th choice we generated, and "[N]" to represent the number of questions we need to generate per term. We use an underline to show we use only one of the choices separated with "/". Additionally, LLMs do not always generate valid outputs. For cases where we do not get enough outputs, we generate multiple times until we obtain enough distinct outputs.

- For procedure "New Term Collection", we use LLMs to get the deduce difficulty of each term. Prompts are detailed in Table 3.
- For procedure "Question and Correct Choice Generation", we use LLMs to get different types of sentences for each of the three tasks. Prompts are detailed in Table 4, Table 5, and Table 6.

- For procedure “Related Term Generation”, we use LLMs to get four different types of related terms for each new term. Prompts are detailed in Table 7, Table 8, and Table 9.
- For procedure “Incorrect Choice Generation”, we only need LLMs to generate incorrect choices for task COMA. Prompts are detailed in Table 10.
- For procedure “LLM Filtering”, we use LLMs to filter all the benchmarks of the three tasks. Prompts are detailed in Table 11, Table 12, and Table 13.
- For evaluation, we follow the prompt of similar datasets in PromptSource [5] to design five prompts manually and select three that have the highest performance for ChatGPT under Gold settings from them. Prompts are detailed in Table 14, Table 15, and Table 16.

Deduce Difficulty	
System Prompt	Please deduce the meaning of the following word based on its spelling, using just one sentence.
User Prompt	What is the meaning of “[W]”? Meaning:

Table 3: Prompt for the Deduce Difficulty.

Sentence Generation	
System Prompt	Please generate [N] different sentences about the new term, each in a separate line, without using the words used above. Make sure that all the sentences you generate have a different subject. Please print the sentence without explanation.
User prompt for COMA	I create a new term “[W]”, which means “[M]”. Please generate [N] different paragraphs about “[W]”, following the format: “Sentence 1. As an effect, / This happened because: Sentence 2.” Sentence 1 should contain “[W]” once. Ensure that it is objective and impartial, focusing on actual actions or events, without any emotional or subjective assumptions. Sentence 2, illustrating the effect / cause of Sentence 1, should be specific to “[W]” in Sentence 1 and not applicable if “[T1]”, “[T2]”, “[T3]”, or “[T4]” is used instead.
User prompt for COST & CSJ	I have created a new term, “[W]”, which means “[M]”. Please generate [N] different sentences about “[W]”, each in a separate line, which should be specific to the meaning of “[W]”. The sentence should be grammatically correct but not applicable if “[T1]”, “[T2]”, “[T3]”, or “[T4]” is used instead.

Table 4: Prompt for the Sentence Generation. We generate N sentences simultaneously for each new term to reduce costs.

Sentence Generation for the Second Half of COST	
System Prompt	Please generate a sentence about the term “[Ti]”, without using the words used above. Make sure that “[Ti]” is exactly in the sentence but not its other forms. Please print the sentence without explanation.
User prompt	Please generate a sentence about “[Ti]”, which should be specific to the meaning of “[Ti]”. The sentence should be grammatically correct but not applicable if “[T1]”, “[T2]”, “[T3]”, or “[T4]” is used instead. Sentences:

Table 5: Prompt for the Sentence Generation for the Second Half of COST. For each generated sentence, we assign different related terms as the answer.

Sentence Generation for the Second Half of CSJ	
System Prompt	Please generate [N] different sentences about the new term, each in a separate line, without using the words used above. Make sure that all the sentences you generate have a different subject. Please print the sentence without explanation.
User prompt	For each sentence generated above, please modify it to use “[W]” illogically, based on the given meaning, while keeping the grammar, fluency, and original subject intact. For each example, print “Wrong Sentence:” and “Corresponding Wrong meaning:” on separate lines, explaining the deviation from the intended meaning. Ensure that each wrong meaning is significantly different from those previously generated.

Table 6: Prompt for the Sentence Generation for the Second Half of CSJ. The user prompt and response of correct sentence generation for CSJ are also used as context input.

Partial Synonym Term Generation	
System Prompt	Please provide three words and three two-word phrases, and display each of them on a separate line. The first three lines are words, each on a separate line, and the last three lines are phrases, each on a separate line. Make sure that there are six lines in total, with each word/phrase at a single line. Do not refrain from answering.
User prompt	Please provide three words and three phrases, “[M]”. Ensure that these are commonly used and easily understood by a 3-year-old child.

Table 7: Prompt for the Partial Synonym Term Generation.

Synonym & Antonym Term Generation	
System Prompt	Please answer the following question by printing three terms without explanation, each at a separate line. If you cannot construct terms that fully meet the requirements, provide terms that partially fulfill the requirements. Do not refrain from answering.
User prompt	What is the synonym / antonym for the new term, “[W]”, that refers to [M]? The synonym / antonym should be a commonly used English term and belong to the same part of speech. Do not use abbreviations and commas, periods in the term, and shorter than five words. Please generate three different alternatives. Synonym / Antonym:

Table 8: Prompt for the Synonym and Antonym Term Generation.

Meaning Guessing Term Generation	
System Prompt	Please answer the following question by printing three terms without explanation, each at a separate line. If you cannot construct terms that fully meet the requirements, provide terms that partially fulfill the requirements. Do not refrain from answering.
User prompt	Please guess the meaning of the term “[W]” and create three alternative terms based on their spelling. Alternative term:

Table 9: Prompt for the Meaning Guessing Term Generation.

COMA Incorrect Choice Generation	
System Prompt	Please generate a sentence with ... words to finish the following paragraph. Please print the sentence without explanation.
User prompt	[Replace the new term [W] in [Question] with its related term [Ti]]. As an effect, / This happened because:

Table 10: Prompt for the COMA Incorrect Choice Generation. For each generated question, we create completions that are correct for each related term as incorrect choices. To make it more challenging to distinguish, we prompt that the lengths of the incorrect choices generated by LLM are as close as possible to the correct ones.

LLM Filtering for COMA	
System Prompt	Please answer the following choice question by selecting the most probable choices. If multiple choices have equal likelihood, you may choose more than one. List the selected choices (A, B, C, D, E, or F) separated by commas.
User prompt	Given that the term “[W]” means “[M]”, please solve the following multiple-choice exercise: Exercise: choose the most plausible alternative. [Question] so / because... A. [C1] B. [C2] C. [C3] D. [C4] E. [C5] F. [C6] Answer:

Table 11: Prompt for the LLM Filtering for COMA.

LLM Filtering for COST	
System Prompt	Please answer the following choice question by selecting the most probable choices. If multiple choices have equal likelihood, you may choose more than one. List the selected choices (A, B, C, D, E, or F) separated by commas.
User prompt	Given that the term “[W]” means “[M]”, please solve the following multiple-choice exercise: [Question] Replace the __ in the above sentence with the correct choice: A. [C1] B. [C2] C. [C3] D. [C4] E. [C5] F. [C6] Answer:

Table 12: Prompt for the LLM Filtering for COST.

LLM Filtering for CSJ	
System Prompt	Please answer the following question with an integer, without any further explanation.
User prompt	Given that “[W]” means “[M]”. On a scale of 0 to 10, with 0 being extremely unlikely and 10 being highly likely, how probable is it that the following sentence is coherent and aligns with general understanding? [Question] Answer:

Table 13: Prompt for the LLM Filtering for CSJ.

COMA Evaluation	
System Prompt for Base Setting	Please answer the following question by printing exactly one choice from “A”, “B”, “C”, “D”, without explanation.
System Prompt for Gold Setting	Given that “[W]” means “[M]”. Please answer the following question by printing exactly one choice from “A”, “B”, “C”, “D”, without explanation.
User prompt 1	Exercise: choose the most plausible alternative. [Question] because / so... A. [C1] B. [C2] C. [C3] D. [C4] Answer:
User prompt 2	[Question] In the previous sentence, does __ refer to A. [C1], B. [C2], C. [C3], or D. [C4]? Answer:
User prompt 3	Fill in the __ in the below sentence: [Question] Choices: A. [C1] B. [C2] C. [C3] D. [C4] Answer:

Table 14: Prompt for the COMA Evaluation.

COST Evaluation	
System Prompt for Base Setting	Please answer the following question by printing exactly one choice from “A”, “B”, “C”, “D”, without explanation.
System Prompt for Gold Setting	Given that “[W]” means “[M]”. Please answer the following question by printing exactly one choice from “A”, “B”, “C”, “D”, without explanation.
User prompt 1	[Question] Replace the __ in the above sentence with the correct choice: A. [C1] B. [C2] C. [C3] D. [C4] Answer:
User prompt 2	[Question] Is this example in line with commonsense and grammatically correct? Answer:
User prompt 3	Given that “[W]” means “[M]”. On a scale of 0 to 10, with 0 being extremely unlikely and 10 being highly likely, how probable is it that the following sentence is coherent and aligns with general understanding? [Question] Answer:

Table 15: Prompt for the COST Evaluation.

CSJ Evaluation	
System Prompt under Base Setting	Please answer the following question by printing “YES / Acceptable” or “NO / Unacceptable”, without explanation.
System Prompt under Gold Setting	Given that “[W]” means “[M]”. Please answer the following question by printing “YES / Acceptable” or “NO / Unacceptable”, without explanation.
User prompt 1	Does the following sentence coherent and aligned with general understanding? Please answer “YES” or “NO”. [Question] Answer:
User prompt 2	[Question] Is this example in line with commonsense and grammatically correct? Answer:
User prompt 3	The following sentence is either “Acceptable”, meaning it fits the commonsense, or “Unacceptable”. Which is it? [Question] Answer:

Table 16: Prompts for the CSJ Evaluation.

C Human Filtering

C.1 Human Filtering Settings

Interactive interface. Our human-interactive interface, built using the SurveyJS library in Vue3 frontend and Flask backend, is designed to provide a user-friendly workflow and efficient annotator experience for our new term benchmark. The platform supports translation, flexible question numbers, and loading history for all three question types. By translating questions into the native language of annotators and providing a clear interface, users can answer questions in about 30 seconds, completing annotations for 900 questions in 10 hours.

Upon accessing the platform, users receive a welcoming message and need to fill in a unique username, ensuring each user can only fill out one questionnaire, as shown in Figure 9. The platform also allows users to decide the total number of questions they wish to answer. The "Loading History" feature enables users to load and modify their previous history. Choosing "Yes" includes all previously answered questions in their total count and allows users to check and change previous answers, while selecting "No" provides new questions.

On the answering page, our interface comprises three question types, as shown in Figure 10. We separate different types of questions into distinct pages, with each page containing 10 questions. Answers are saved after annotators finish any page, making it easy for them to skip and return to continue at any time. Finally, to support situations with no choices and to provide feedback and records for special cases, we have set up two additional choices, namely "None" and "Other".

Annotators. For human filtering, we recruited two crowdsource annotators and one professional annotator. For the crowdsource annotators, we enlisted the services of two English-proficient annotators from China via a crowdsourcing platform. After evaluation, we determined the annotation cost to be RMB 1.5 per question per person. For the professional annotator, we engaged a university professional annotator, who is a current master's student specializing in natural language processing, to perform the annotation.

To minimize inconsistencies, we provide users with detailed guidance, including annotation instructions, examples, and requirements. Specifically, for multiple-choice questions, annotators are asked to select the choice that best aligns with the question's intent. If multiple choices have similar probabilities and are all reasonable, they should select multiple choices. If none of the choices are reasonable, they should choose "None". Based on our evaluation and filtering experience with LLMs on NewTerm, we observed that these annotation criteria closely resemble the standards used for most LLMs. Since our benchmark aims to evaluate the performance of LLMs, we chose criteria for human annotation that align as closely as possible with LLMs.

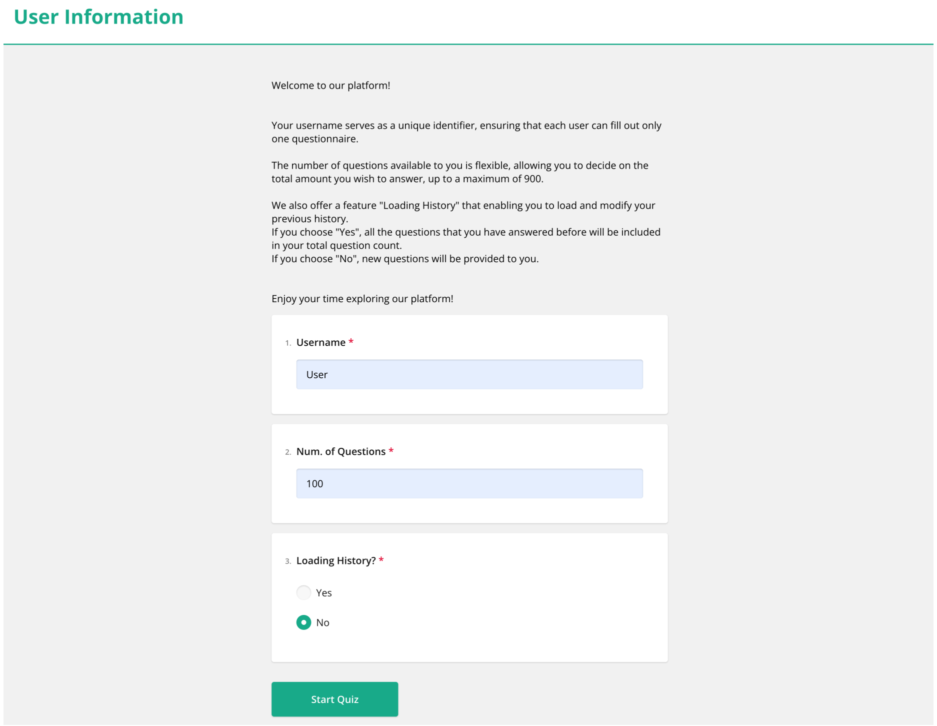
Additionally, to increase efficiency and reduce annotation costs, we provide translations of the questions. To minimize bias introduced by translation, we require annotators to be proficient in English during the recruitment process. We also emphasize in our instructions that translations may be inaccurate due to the presence of new terms and ask annotators to use translations only for supplementary understanding while basing decisions solely on the English question. Our final decision is made by the professional annotator with strong English reading and writing skills, who can better adhere to our requirements. This approach helps minimize potential risks of errors and ambiguities while achieving lower annotation costs and higher annotation efficiency.

C.2 Analysis of Human Filtering

Filtering reason analysis. We analyze the reasons for answers that do not align with the three human annotations under NewTerm 2022 and 2023, i.e., humans choosing more than one choice (Multi.), no choices (Zero), or choosing choices differing from auto-generated ones (Wrong). Results are in Table 17. In our construction pipeline, "Multi." is caused by LLM filtering, which failed to choose all the incorrect

	Multi.	Zero	Wrong	Acc. (%)
COMA	102	112	202	76.89
COST	129	142	77	80.67
CSJ	-	-	281	84.39

Table 17: The number of cases where the automatically generated answer does not align with human annotation. "Acc." denotes the percentage of non-alignments, with "Multi.", "Zero", and "Wrong" denotes the number of errors defined in Appendix C.2.



User Information

Welcome to our platform!

Your username serves as a unique identifier, ensuring that each user can fill out only one questionnaire.

The number of questions available to you is flexible, allowing you to decide on the total amount you wish to answer, up to a maximum of 900.

We also offer a feature "Loading History" that enabling you to load and modify your previous history.
If you choose "Yes", all the questions that you have answered before will be included in your total question count.
If you choose "No", new questions will be provided to you.

Enjoy your time exploring our platform!

1. Username *

User

2. Num. of Questions *

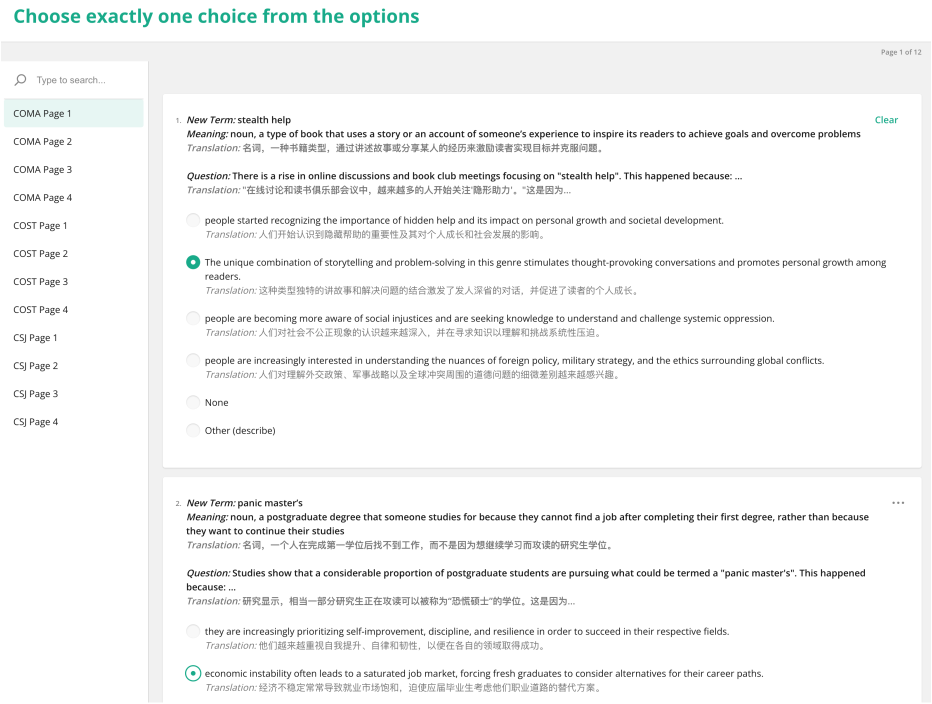
100

3. Loading History? *

○ Yes

● No

Figure 9: Welcome page of the human-interactive interface, displaying a welcoming message and choices for loading history and selecting the number of questions.



Choose exactly one choice from the options

1. ***New Term:*** **stealth help**
Meaning: **noun, a type of book that uses a story or an account of someone's experience to inspire its readers to achieve goals and overcome problems**
Translation: 名词，一种书籍类型，通过讲述故事或分享某人的经历来激励读者实现目标并克服问题。

Question: **There is a rise in online discussions and book club meetings focusing on "stealth help". This happened because: ...**
Translation: "在线讨论和读书俱乐部会议中，越来越多的人开始关注'隐形助力'。"这是因为...

○ people started recognizing the importance of hidden help and its impact on personal growth and societal development.
Translation: 人们开始认识到隐藏帮助的重要性及其对个人成长和社会发展的影响。

● The unique combination of storytelling and problem-solving in this genre stimulates thought-provoking conversations and promotes personal growth among readers.
Translation: 这种类型独特的讲故事和解决问题的结合激发了发人深省的对话，并促进了读者的个人成长。

○ people are becoming more aware of social injustices and are seeking knowledge to understand and challenge systemic oppression.
Translation: 人们对社会不公正现象的认识越来越深入，并在寻求知识以理解和挑战系统性压迫。

○ people are increasingly interested in understanding the nuances of foreign policy, military strategy, and the ethics surrounding global conflicts.
Translation: 人们对理解外交政策、军事战略以及全球冲突周围的道德问题的细微差别越来越感兴趣。

○ None

○ Other (describe)

2. ***New Term:*** **panic master's**
Meaning: **noun, a postgraduate degree that someone studies for because they cannot find a job after completing their first degree, rather than because they want to continue their studies**
Translation: 名词，一个人在完成第一学位后找不到工作，而不是因为想继续学习而攻读的研究生学位。

Question: **Studies show that a considerable proportion of postgraduate students are pursuing what could be termed a "panic master's". This happened because: ...**
Translation: 研究显示，相当一部分研究生正在攻读可以被称为"恐慌硕士"的学位。这是因为...

○ they are increasingly prioritizing self-improvement, discipline, and resilience in order to succeed in their respective fields.
Translation: 他们越来越重视自我提升、自律和韧性，以便在各自的领域取得成功。

● economic instability often leads to a saturated job market, forcing fresh graduates to consider alternatives for their career paths.
Translation: 经济不稳定常常导致就业市场饱和，迫使应届毕业生考虑他们职业道路的替代方案。

Figure 10: Answering page of the human-interactive interface, showcasing the three tasks in NewTerm benchmark: COMA, COST, and CSJ.

choices that are reasonable, covering 22.11% of the incorrect cases. “Zero” is caused by question generation, where LLMs do not understand the new term correctly and generate meaningless questions or incorrect answers. All errors in CSJ are also caused by this reason. It covers 51.19% of the cases. “Wrong” means that both are partly incorrect; the correct answer is not entirely correct, and LLMs fail to choose all choices that are more plausible than the correct answer. This covers 26.70% of the cases. Stronger LLMs may further alleviate this problem and make the pipeline more reliable.

Subprocess analysis. As a cascaded generation benchmark, error propagation can often occur between subprocesses, making it necessary to analyze the error rates for each subprocess. In our framework, there are two types of error propagations in these steps:

- First, the results of “Related Term Generation” are used for “Incorrect Choice Generation” of COMA and COST questions with answers being old terms. However, these COST questions aim to generate fill-in-the-blank questions related to old terms. As long as a valid term is generated, valid questions can still be generated.

- Second, the results of “Question and Correct Choice Generation” are used for “Incorrect Choice Generation” of COMA and CSJ questions with the answer “False”. However, these CSJ questions aim to generate incorrect sentences in the judgment task. Even if the first part of the sentence is not correct, valid incorrect sentences can still be generated.

Therefore, the error propagation problem mainly occurs in the generation of the COMA dataset. To further quantitatively assess the impact of error propagation problems, we randomly select 50 cases of the COMA task in NewTerm 2022 for human annotation. The “Related Term Generation” procedure has a 7.60% error probability, where the generated term is less related to the new term. The “Question and Correct Choice Generation” procedure has a 12.00% error probability, where the generated sentence is incorrect for the new term.

The “Incorrect Choice Generation” procedure is based on the output of the above procedures. Additionally, incorrect questions should be discarded regardless of the choices, so we ignore cases with incorrect questions in the subsequent annotation process. Two types of errors occur in incorrect choice generation: 1) First, the generated incorrect choices are reasonable under the current question, covering 26.89% of choices. 2) Second, due to error propagation from the related term, the choice may be irrelevant to the original question. However, we did not observe this phenomenon in the 264 annotated choices with valid questions. This is because the question occupies the main part of the prompt, and a single irrelevant term is not enough to interfere with LLMs to generate irrelevant choices.

LLM	Size	NewTerm 2022 w/ human filtering					NewTerm 2022 w/o human filtering				
		COMA	COST	CSJ	Avg.	Gold	COMA	COST	CSJ	Avg.	Gold
Llama-2-Chat	7B	28.89	28.12	60.88	39.29	58.68	31.56	28.89	58.67	39.70	56.33
	13B	31.24	33.19	56.11	40.18	60.92	30.78	33.56	57.11	40.48	58.67
	70B	45.49	48.99	61.13	51.87	82.38	45.11	51.33	61.67	52.70	78.48
Llama-3-Instruct	8B	52.94	46.81	63.19	54.31	88.19	51.67	51.67	63.78	55.70	85.41
	70B	66.01	58.70	66.15	63.62	96.07	66.78	62.33	67.00	65.37	94.85
Claude-Instant-1.2	S	49.28	47.54	68.60	55.14	88.33	49.56	52.00	68.22	56.59	86.22
Claude-2.1	M	38.04	54.20	71.94	54.73	82.20	37.89	56.44	70.67	55.00	79.22
Claude-3-haiku	S	58.04	53.62	67.18	59.61	92.60	58.89	57.67	68.00	61.52	90.56
Claude-3-sonnet	M	56.73	56.23	64.48	59.15	93.73	56.22	58.33	65.56	60.04	92.19
Claude-3-opus	L	64.58	67.97	65.38	65.98	93.60	64.78	70.00	67.00	67.26	92.85
GPT-3.5-0613	S	52.42	49.71	73.62	58.58	87.71	52.89	53.56	72.67	59.70	85.30
GPT-3.5-0125	S	51.37	49.86	72.07	57.77	87.63	52.56	54.44	71.33	59.44	84.78
GPT-4-0613	L	68.37	61.16	70.14	66.56	98.91	70.78	65.22	70.33	68.78	98.59
GPT-4-1106	M	72.03	63.48	70.79	68.76	97.56	71.78	67.22	71.11	70.04	97.11
GPT-4-0125	M	69.80	65.94	71.94	69.23	98.11	70.33	68.78	72.56	70.56	97.70
Average	-	53.68	52.37	66.91	57.65	87.11	54.11	55.43	67.05	58.86	85.22

Table 18: Results for different LLMs under benchmark with and without human filtering. The definitions of abbreviation are identical with Table 1.

Evaluation result difference before and after human filtering. We compared the test results of different LLMs under NewTerms 2022, both before human filtering (900 questions) and after human filtering (744 questions). The results are shown in Table 18. We can see that the results under the two benchmark settings are highly consistent. The absolute value of the performance gap between the two settings averages only 1.59 across each different task and each different LLM. Furthermore, the performance ranking among different models remains entirely consistent under both the Base and Gold settings. This proves that our benchmark can achieve the same evaluation abilities and conclusions without human filtering, indicating that human filtering is optional.

Additionally, for the filtered-out questions, the performance of LLMs is slightly higher under the Base setting (+0.95 on average) but lower under the Gold setting (-1.89 on average) compared to the unfiltered questions. This suggests that these filtered-out questions may be biased towards LLMs, making it easier for them to select the auto-generated answers, even though the questions themselves may not be correct.

D Main Results on More Open-Sourced LLMs

We also employ the following LLMs for our experiments: Vicuna-1.3 (7B and 13B) [77], fine-tuned from Llama [61]; ChatGLM-2 (6B) [74]; Baichuan-2 (7B and 13B) [69]; Qwen (7B and 14B) [6]; and Mistral (7B) [35]. All tests are done under greedy decoding. Experimental results are shown in Table 19. As indicated by the results, except for Vicuna-1.3, which performed poorly on our tasks and failed to understand the question well, the experimental results of the remaining models all maintain the conclusions obtained in the main text. Among them, the Qwen-Chat model achieved the best results on both Base and Gold, followed by Mistral-Instruct-0.1.

LLM	Size	NewTerm 2022					NewTerm 2023				
		COMA	COST	CSJ	Avg.	Gold	COMA	COST	CSJ	Avg.	Gold
Vicuna-1.3	7B	30.46	24.78	58.94	38.06	44.20	25.88	32.77	71.85	43.50	44.49
	13B	30.59	23.91	65.77	40.09	43.73	25.88	32.34	80.08	46.10	50.02
ChatGLM-2	6B	42.09	43.77	51.99	45.95	64.43	31.87	60.17	56.31	49.45	62.07
Baichuan-2-Chat	7B	40.00	42.90	63.06	48.65	72.90	48.25	58.05	79.02	61.77	74.72
	13B	41.44	50.72	60.10	50.76	76.88	46.78	64.55	64.67	58.67	76.71
Qwen-Chat	7B	44.31	50.43	68.08	54.28	83.95	46.35	65.11	83.53	65.00	85.22
	14B	50.85	49.13	68.73	56.24	87.14	56.43	65.54	83.13	68.37	90.10
Mistral-Instruct-0.1	7B	43.53	42.61	56.76	47.63	79.25	44.44	57.49	66.80	56.24	80.31
Average	-	40.41	41.03	61.68	47.71	69.06	40.75	54.50	73.17	56.14	70.46

Table 19: Results for more different LLMs on NewTerm 2022 and 2023. The order of the LLMs is based on their release date in HuggingFace, with the earliest at the top. The definitions of the abbreviations are the same as in Table 1.

E Case Study for LLMs of Different Year

We also present specific examples that illustrate the differences in how earlier models and more recent models interpret new terms, highlighting the advancements made by newer models in understanding recent or domain-specific vocabulary. To further explore this, we analyzed cases involving Llama-2-Chat-70B and Llama-3-Instruct-70B, focusing on concepts that earlier LLMs overlooked but more recent models successfully identified.

- **New Term:** ***supercloud***
- **Meaning:** Noun, a single computing system where services such as storage, apps, etc. from different providers can be easily accessed by the user.
- **Question:** Businesses are adopting ***superclouds*** to streamline integration across various digital service platforms. Is this example in line with commonsense and grammatically correct?
- **Llama-2 Response:** Incorrect (X)
- **Llama-3 Response:** Correct (✓)
- **Llama-2 Meaning Guessing:** A supercloud is a massive, powerful cloud that is formed by the combination of several smaller clouds, suggesting a large and potentially threatening weather system.
- **Llama-3 Meaning Guessing:** The word "supercloud" likely refers to an extremely large or powerful cloud, either in a literal sense (e.g., a massive storm cloud) or a figurative sense (e.g., a vast and dominant cloud computing platform).

In response to our question containing the new term "supercloud," under the zero-shot Base setting, Llama-2 incorrectly labeled this as "Incorrect," whereas Llama-3 accurately classified it as "Correct." To further investigate, we analyzed how each model interpreted the meaning of the term. We found that Llama-2 solely associated the term with meteorological contexts, while Llama-3 correctly connected it to cloud computing. This difference highlights the older model's limitations and misjudgments due to its incomplete grasp of newer technological terms.

Additionally, we present another case study that explores different types of new terms and tasks:

- **New Term:** ***stochastic parrot***
- **Meaning:** Noun, a way of describing a large language model, because it can produce text that sounds natural but does not understand what it is saying.
- **Question:** The _ flawlessly recites poetry without grasping the underlying emotions. In the previous sentence, does _ refer to A. ***Stochastic parrot***, B. Aware person, C. Probabilistic repeater, or D. Stocky patriot?
- **Llama-2 Response:** C (X)
- **Llama-3 Response:** A (✓)
- **Llama-2 Meaning Guessing:** A stochastic parrot is a parrot that engages in random and unpredictable behavior, possibly due to its exposure to certain environmental factors or its natural temperament.
- **Llama-3 Meaning Guessing:** The term "stochastic parrot" likely refers to a machine learning model or artificial intelligence that generates responses or outputs in a seemingly random or unpredictable manner, much like a parrot mimicking sounds, but with a nod to the mathematical concept of stochasticity, implying a probabilistic or chance-based process.

This case demonstrates that Llama-2 perceived the term "stochastic parrot" in its literal sense, leading to a misinterpretation of the task, while Llama-3 accurately recognized its metaphorical usage to describe an AI's capabilities, correctly guiding its response to the question.

F Benchmark Construction with Different LLMs

In the main text, we primarily used `gpt-4-0613` to generate the benchmark. It is worth noting that although our benchmark generation process benefits from stronger LLMs, it does not rely on any specific LLM. As for the universality of our pipeline with other LLMs, we constructed a new benchmark using Claude, i.e., `claude-2.1`, based on the 300 new words we collected in 2022. We also employed the same construction framework and filtering methods. Finally, before human filtering, we obtained 900 questions, aligning with the generation of NewTerm 2022.

	NewTerm 2022 with GPT-4				**NewTerm 2022 with Claude-2.1**			
	Multi.	**Zero**	**Wrong**	**Acc. (%)**	**Multi.**	**Zero**	**Wrong**	**Acc. (%)**
COMA	49	54	97	77.78	81	51	176	65.78
COST	50	55	30	85.00	36	136	38	76.67
CSJ	-	-	140	84.44	-	-	121	86.56

Table 20: The number of cases where the automatically generated answer does not align with human annotation. The abbreviations are the same as defined in Table 17.

Subsequently, we adopted the same human filtering approach as the main text. We calculate the inter-annotator agreement using Fleiss' Kappa, which reaches a score of 0.67. Additionally, in 76.33% of cases, the annotator results match the automatically generated ones. These results are slightly lower than those of GPT-4 (0.70 / 82.41%), but still comparable. Detailed analysis of error reasons is given in Table 20. For the COMA task, which requires a multi-step generation process, the error rate is more significantly affected by the LLM's capabilities. However, for tasks that only require one or two steps of generation, such as CSJ, the impact is smaller.

We further analyze the performance of different LLMs under NewTerm 2022, with experimental settings aligned with those in the main text in Section 4.1. The results are shown in Table 21. The performance ranking of different LLMs and the performance changes under different settings are consistent with NewTerm 2022 generated by `gpt-4-0613`, demonstrating the effectiveness of using different LLMs to generate benchmarks.

LLM	**Size**	**Base**				**Gold**			
		COMA	**COST**	**CSJ**	**Avg.**	**COMA**	**COST**	**CSJ**	**Avg.**
Vicuna-1.3	7B	31.07	26.50	57.92	38.49	32.85	29.43	69.50	43.92
	13B	35.71	27.75	61.00	41.49	31.88	29.15	86.87	49.30
ChatGLM-2	6B	46.44	46.44	46.46	46.45	74.92	73.92	46.98	65.27
Llama-2-Chat	7B	31.55	25.80	78.12	45.16	55.34	61.09	89.96	68.80
	13B	46.12	34.45	33.72	38.10	72.98	54.53	48.91	58.81
	70B	56.15	43.65	41.96	47.25	88.35	78.94	74.77	80.69
Baichuan-2	7B	52.27	44.49	72.97	56.58	77.51	74.06	64.99	72.19
	13B	57.61	45.75	55.73	53.03	79.94	76.57	66.02	74.18
Qwen	7B	53.07	47.70	63.58	54.78	80.58	76.29	77.48	78.12
	14B	57.77	45.19	74.13	59.03	88.03	90.10	89.19	89.10
Mistral	7B	49.68	44.77	44.92	46.45	84.63	77.27	57.53	73.14
Llama-3-Instruct	8B	61.49	47.56	52.64	53.90	91.59	89.12	90.09	90.27
	70B	68.12	60.11	53.02	60.42	95.31	96.37	88.55	93.41
GPT-3.5-0613	-	61.00	48.54	67.44	58.99	88.35	91.21	89.96	89.84
GPT-4-0613	-	71.04	60.81	59.72	63.85	94.98	97.07	91.63	94.56
Average	-	51.94	43.30	57.56	50.93	75.82	73.01	75.50	74.77

Table 21: Results for different LLMs on benchmark generated by `claude-2.1` based on terms from 2022. The order of the LLMs is based on their release date in HuggingFace, with the earliest at the top, except for GPT series models. The definitions of abbreviation are identical with Table 1.

G Datasheet for NewTerm

In this section, we provide more detailed documentation of the dataset with the intended uses. We base ourselves on the datasheet proposed by Gebru et al. [24].

G.1 Motivation

For what purpose was the dataset created? The NewTerm benchmark focuses on the real-time evaluation of LLMs, which is crucial for their effectiveness. Specifically, we concentrate on the less-explored area of new term evaluation and propose a highly automated benchmark construction pipeline to ensure real-time updates and generalization to a wider variety of terms. Our ultimate goal is to develop an efficient benchmark for tracking LLMs' ability to understand new terms, and we will update it annually. Furthermore, we can also assess the performance of different LLMs and potential improvement strategies.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? The NewTerm benchmark was developed with contributions from the authors of this paper and was supported by the Institute of Computing and Intelligence at Harbin Institute of Technology, Shenzhen, China.

Who funded the creation of the dataset? The dataset was funded by multiple grants, as detailed in the acknowledgments section.

G.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Each instance consists of a question covering three tasks, introduced in Section 3.3. These questions are generated in a highly automated manner by our construction pipeline.

How many instances are there in total (of each type, if appropriate)? The benchmark currently consists of 744 questions for NewTerm 2022, and 715 for NewTerm 2023, evaluating the performance of LLMs under in total 600 new terms. We will update the benchmark annually to evaluate the latest year's new terms.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? The NewTerm benchmark is a sample of instances from a larger set, where the large set corresponds to the benchmark composed of questions for all new terms collected annually in online dictionaries. We select the most representative 300 new terms from the full set of updated terms each year, covering new words, new phrases, and old words with new meanings, and construct benchmarks for these new terms. This sample covers the most challenging part of the annual new term updates and serves as a typical representation of the full set. For a detailed analysis, please refer to Section 4.4.

What data does each instance consist of? For all tasks, each instance is given in JSON format, including the evaluated "new term", its "meaning", and its "type" by this question. Here, new words correspond to the type "new words not deduced", new phrases correspond to "new phrases not deduced", and old words with new meanings correspond to "old words not deduced". Additionally, it includes a "question", two or four "choices", and the correct answer "gold", which represents the index of the correct choice. For COMA, we additionally include a "split" attribute, indicating whether the selected choice is the cause or the effect of the question. This will correspond to different testing prompts. Below is an example from the COMA task:

```
{
    "term": "Juggers",
    "meaning": "When the sleeves of a shirt are uncomfortably
       short.",
    "type": "new words not deduced",
    "question": "Several people have started complaining about
       their new Juggers.",
    "choices": [
        "the company had used low-quality materials, leading to
            rapid wear and tear, much to the customers’
           disappointment and dissatisfaction.",
        "the company failed to clearly communicate the product’
           s dimensions, leading to widespread frustration
           among their customer base.",
        "the fabric quality was sub-par, colors faded after a
           few washes, and sizes were not accurately
           represented on the website.",
        "the trend of body-hugging shirts has led to a spate of
            situations where people ended up with sleeves
           shorter than preferred."
    ],
    "gold": 3,
    "split": "cause"
}
```

Is there a label or target associated with each instance? Yes, as mentioned in the previous question, each instance includes a “gold” field, which corresponds to the index of the correct answer choice.

Is any information missing from individual instances? No, all the instances should have complete information corresponding to the content as well as to the attributes.

Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)? For each selected new term, we construct multiple instances covering various tasks to evaluate LLMs’ understanding ability. To make this relationship explicit, we can match the “term” and “meaning” fields in the instances. Instances with identical term and meaning fields indicate that they are evaluating the same new term.

Are there recommended data splits (e.g., training, development/validation, testing)? The NewTerm benchmark primarily focuses on evaluation, and all instances are part of the test set. For training, we recommend using only the “term” and “meaning” fields in each instance. A clean dataset containing only these two fields is also released and can be directly accessed.

Are there any errors, sources of noise, or redundancies in the dataset? Before human filtering, the NewTerm benchmark contains errors and sources of noise, which are analyzed in detail in Appendix C.1. After human filtering, we effectively removed these errors and noise. There are no redundancies in our benchmark.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? The NewTerm benchmark is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals’ non-public communications)? No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? No.

G.3 Collection Process

How was the data associated with each instance acquired? We initially collected new terms from online dictionaries, including Cambridge[2], Collins[3], and Oxford[4]. Subsequently, the NewTerm benchmark was indirectly derived from other data using our automated framework, as detailed in Section 3.4. We validated and filtered the generated data through human filtering and thorough analysis, as described in Section 3.5.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? We downloaded the HTML of online dictionary update pages and extracted new terms and their meanings, which typically correspond to fixed fields. Due to the neat and noise-free format of the dictionaries, we did not need to perform further filtering or validation.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? The sampling method is described in Section 3.2, and its further verification can be found in Section 4.4.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? Please refer to Appendix C.1.

Over what timeframe was the data collected? Our benchmark is related to the new terms of each year. Currently, NewTerm 2022 covers new terms from January 2022 to March 2023, and NewTerm 2023 covers April 2023 to March 2024. Additionally, we plan to update the benchmark annually, covering new terms from April of each year to March of the following year.

Were any ethical review processes conducted (e.g., by an institutional review board)? N/A.

G.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? Yes. See Section 3.5.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? Yes. Both the raw and filtered datasets have been released and can be accessed at `https://github.com/hexuandeng/NewTerm`. The filtered datasets are distinguished by the suffix "_clean".

Is the software that was used to preprocess/clean/label the data available? Yes. We have released the automatic pipeline code for LLM filtering, along with all corresponding frontend and backend codes required for human filtering. These can be accessed at the above url.

G.5 Uses

Has the dataset been used for any tasks already? Yes. In our submitted paper, we conducted extensive evaluation and comprehensive analysis on numerous versions of mainstream LLMs, aiming to evaluate their performance when facing new terms, as detailed in Section 4.

Is there a repository that links to any or all papers or systems that use the dataset? N/A.

[2] `https://dictionaryblog.cambridge.org/category/new-words`

[3] `https://www.collinsdictionary.com/submissions/latest`

[4] `https://www.oed.com/information/updates`

What (other) tasks could the dataset be used for? The NewTerm benchmark can also be used for evaluating the performance of LLMs on various other terms beyond new ones, such as religious, literary, and low-frequency terms. To facilitate this, we have released the code for automatic benchmark construction and the human interactive interface construction. This enables developers interested in building their benchmarks for other new terms to do so with ease. Our construction solution is cost-effective, especially when the human filtering step is omitted, making it accessible for developers to build their own benchmarks. We hope this contribution will encourage further research on the performance of different types of terms within the research community.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? No.

Are there tasks for which the dataset should not be used? No.

G.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? Yes, the dataset is of public access.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? The NewTerm benchmark will be made public on a GitHub repository, which can be found at `https://github.com/hexuandeng/NewTerm`. The public content includes the following three parts:

- NewTerm benchmark: Currently, it covers NewTerm 2022 and NewTerm 2023, constructed from new terms in 2022 and 2023, and will continue to be updated annually.
- Testing code: We have released easy-to-use testing code and corresponding instructions, allowing testing on most open-source/closed-source LLMs with just a few commands. For testing other LLMs, we provide detailed guidance, enabling developers to modify minimal code to test their LLMs. Finally, all results in this paper are consistent with the testing framework, ensuring the reproducibility of the reported results.
- Benchmark construction code: We have released the code for automatic benchmark construction and human interactive interface. This supports developers interested in building their benchmarks for other new terms, e.g., religious, literary, and low-frequency terms.

When will the dataset be distributed? The NewTerm benchmark is currently available in the GitHub repository referenced in the previous response.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? The NewTerm benchmark is distributed under a Creative Commons Attribution 4.0 International license (CC BY 4.0).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? No.

G.7 Maintenance

Who will be supporting/hosting/maintaining the dataset? The maintenance and extension of NewTerm will be carried out by the authors of the paper.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? For inquiries, please contact hxuandeng@gmail.com.

Is there an erratum? No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? Yes, we will update the benchmark annually to evaluate the performance of the newest LLMs under new terms from the most recent year, covering the period from April of the current year to March of the following year. The authors of this paper will collect these new terms, construct the updated benchmark, and release it on the GitHub repository mentioned in the previous question.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? N/A.

Will older versions of the dataset continue to be supported/hosted/maintained? Yes, we will continue to support, host, and maintain older versions of the dataset in the open-source repository. This will enable tracking the performance of LLMs over time as terms evolve.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? Yes. We have released the code for automatic benchmark construction and the human interactive interface construction, which supports developers interested in building their benchmarks for other new terms. Contributors can use these codes to generate datasets for model evaluation or improvement. The new datasets can be distributed independently by the contributors themselves, or they can contact the authors of this paper via email. We will manually review them and decide whether to publish them in the GitHub repository.

G.8 Further Statement

- The authors of the paper bear all responsibility in case of violation of rights, etc., and confirmation of the data license. We confirm the use of the CC BY 4.0 license for the data.
- We ensure that all results are easily reproducible in Appendix G.6, guarantee that all results can be easily reproduced, i.e. all necessary datasets, code, and evaluation procedures are accessible and documented in our GitHub repository.
- We release the NewTerm benchmark along with the associated construction and evaluation code at `https://github.com/hexuandeng/NewTerm`, ensuring that the dataset will be available for a long time. We will continue hosting and maintaining this benchmark, updating it annually with the latest year's data to support tracking the real-time abilities of LLMs. The dataset format is in JSONL.
- To ensure our benchmark can be discovered and organized by anyone, we publish it on Hugging Face at `https://huggingface.co/datasets/hexuandeng/NewTerm`, which will automatically add structured metadata to the dataset.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 1. We introduces NewTerm, an annually updating benchmark for tracking the performance LLMs on new terms. Results and further analysis reveal the characteristics and reasons for terms that pose challenges to LLMs, facilitating future research.
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [No] Our work does not have potential negative societal impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We have read the ethics review guidelines and ensured that our paper has no risks associated with the proposed data collection and data usage.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Abstract and Appendix G.6.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.1.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We use greedy search, and the models will provide deterministic results.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.1.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We cite every model we used in our work, detailed in Section 4.1.
   (b) Did you mention the license of the assets? [No] All the assets we used in our work are licensed for research use.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See Section 4.1 and Appendix D.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] All the data we used in our work are licensed for research use.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] We only use online dictionaries as input and employ LLMs that have undergone safety alignment for output. The data we are using and curating does not contain personally identifiable information or offensive content.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See Appendix C.1.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [No] We only ask annotators to answer multiple-choice questions, and the questions do not contain any offensive content.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] See Appendix C.1.