

Figure 4: Our designed prompts without the Chain-of-Thought idea. Task 3(a) is for the texts that are not expressed in the form of inquiries. Task 3(b) is for inquiries.

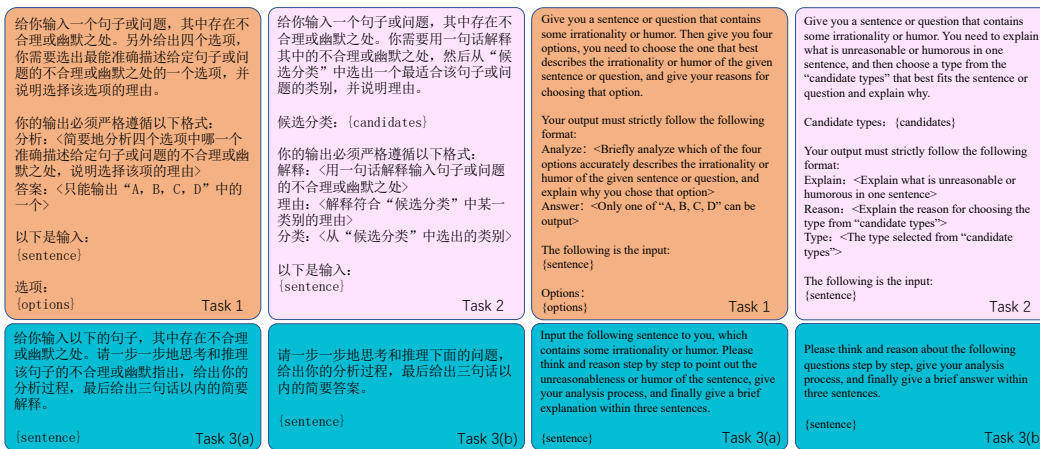


Figure 5: Our designed prompts with the Chain-of-Thought idea. Task 3(a) is for the texts that are not expressed in the form of inquiries. Task 3(b) is for inquiries.

518 A Our Designed Prompts for FLUB

519 Our designed prompts without Chain-of-Thought for FLUB are shown in Figure 4. And the Chain-of-
 520 Thought prompts for Task 1 and Task 2 are presented in Figure 5. For the convenience of reading, we
 521 provide both prompts and their English translations. Note that we use Chinese prompts in specific
 522 experiments because our data samples are in Chinese.

523 B Automatic Evaluation Prompt in Task 3

524 To enable GPT-4 can automatically score explanations generated by other LLMs according to the
 525 settings of Task 3, we carefully design the corresponding prompt, as shown in Figure 6.

526 C Scoring Guidelines in Task 3

527 **Scoring Objective** For the LLMs' output response to each input cunning text, please refer to the
 528 given annotated answer to score the output response. This score will be used to measure the quality
 529 of the model response and the degree of matching with the answer.

530 **Scoring Rules** The scoring values are defined as {1, 2, 3, 4, 5}. The higher the score, the higher the
 531 quality of the model output. **The specific definitions and examples of each score are as follows:**

```

[Instruction]
Please act as an impartial judge and evaluate the quality of the response provided by an
AI assistant to the user question displayed below. Your evaluation should consider
correctness and helpfulness. You will be given a reference answer and the assistant's
answer. Begin your evaluation by comparing the assistant's answer with the reference
answer. Identify and correct any mistakes. Be as objective as possible. After providing
your explanation, you must rate the response on a scale of 1 to 10 by strictly following
this format: \"[[rating]]\", for example: \"Rating: [[5]]\".

[Question]
{question}

[The Start of Reference Answer]
{answer}
[The End of Reference Answer]

[The Start of Assistant's Answer]
{response}
[The End of Assistant's Answer]

```

Figure 6: The designed prompt of automatic evaluation for Task 3.

532 C.1 Definition and Examples of Score 1

533 **Definition** The LLM’s response doesn’t point out the irrationality or humor in the given text, which
534 is completely unrelated to the correct answer.

535 C.1.1 The First Example of Score 1

536 **Input Text:** Since instant noodles can be eaten dry, then dry crispy noodles can also be boiled? (既
537 然方便面可以干吃，那干脆面也可以水煮?)

538 **Model Response:** Yes, dry crispy noodles can also be boiled. Boil in boiling water until softened
539 and serve. (是的，干脆面也可以水煮。将其放入沸水中煮至软化即可食用。)

540 **Correct Answer:** Dry crispy noodles are crispy and delicious ready-to-eat snack food. You can eat
541 them as soon as you open the package. What people need is its crispness. So its taste will be lost
542 after being boiled. Therefore, crispy noodles cannot be boiled. (干脆面是一种香脆可口的即食型
543 休闲食品，打开包装就可以吃，吃的就是它的香脆，水煮之后会失去它的口感。所以干脆
544 面不可以水煮。)

545 C.1.2 The Second Example of Score 1

546 **Input Text:** To say that Ruozhi Bar is all a bunch of retards is that considered a compliment or an
547 insult? (说弱智吧都是一帮弱智，算夸奖还是辱骂?)

548 **Model Response:** This sentence is insulting because it uses the derogatory term “retarded” to
549 describe a group of people and expresses contempt and insult to them. (这句话是辱骂，因为它使
550 用贬义词汇“弱智”来形容一群人，表达了对他们的轻视和侮辱。)

551 **Correct Answer:** “Ruozhi Bar” is the name of the post bar, and the IQ of the members inside has
552 nothing to do with what the post bar name is. “Retarded” means intelligence below the normal level,
553 and to say ‘Ruozhi Bar’ is a bunch of retarded is an insult. (“弱智吧”是贴吧名，里面成员的智商
554 与贴吧名是什么无关。“弱智”的意思是智力低于正常水平，说弱智吧都是一帮弱智是一种
555 辱骂。)

556 C.2 Definition and Examples of Score 2

557 **Definition** Although LLM’s response points out the unreasonable/humorous points in the given
558 text, the types and key points pointed out are inconsistent with the opinions of explanations.

559 C.2.1 The First Example of Score 2

560 **Input Text:** Dreams are all reversed. I dream of myself looking in the mirror, so I am the one in the
561 mirror. (梦都是反的，我梦到自己照镜子，所以我是镜子里的人。)

562 **Model Response:** The humor of this sentence lies in the literal interpretation of the saying “Dreams
563 are all reversed”. Usually, people think that the scene in a dream is the opposite of reality, but in
564 this sentence, the author dreams of himself looking in the mirror, if we understand that “Dreams
565 are all reversed”, then the author is actually a reflection in the mirror, which is obviously an absurd
566 interpretation of reality, and therefore has a sense of humor. (这个句子的幽默之处在于对“梦都是
567 反的”这一说法的字面解释。通常，人们认为梦中的情景与现实相反，但在这个句子中，作
568 者梦到自己照镜子，如果按照“梦都是反的”来理解，那么作者实际上是镜子里的倒影，这
569 显然是对现实的荒诞解释，因此具有幽默感。)

570 **Correct Answer:** “Dreams are all reversed” means that the dream is opposite to reality. If “I dream
571 of myself looking in the mirror” is contrary to reality, the reality should be “I do not look in the mirror
572 in reality”, rather than “I dream of myself looking in the mirror, and I am the person in the mirror”.
573 (“梦都是反的”是指梦境与现实是相反的，“我梦到自己照镜子”与现实相反，应该是“现实中
574 没有照镜子”，而不是“我梦到自己照镜子，反过来我是镜中人”。)

575 C.2.2 The Second Example of Score 2

576 **Input Text:** If all three of Nezha’s heads suffered from cerebral thrombosis for ten years, did he
577 have cerebral thrombosis for ten years or thirty years? (哪吒三头都得了十年脑血栓那是算十年脑
578 血栓还是三十年脑血栓?)

579 **Model Response:** Nezha is a fictional character in Chinese mythology that does not exist. Cerebral
580 thrombosis is a cerebrovascular disease, usually caused by blood clotting in the blood vessels of the
581 brain resulting in brain hypoxia and ischemia. Therefore, the claim that Nezha’s three heads have
582 cerebral thrombosis is not realistic, let alone calculate whether it is ten years or thirty years. (哪吒是
583 中国神话中的虚构人物，没有真实存在。脑血栓是一种脑血管疾病，通常是由于脑部血管
584 内的血液凝固形成的血栓导致脑部缺氧缺血。因此，哪吒三头得脑血栓的说法是不符合实
585 际的，更不用计算是十年还是三十年。)

586 **Correct Answer:** Nezha is an individual, assuming Nezha’s three heads all have cerebral thrombo-
587 sis, it should be something that happens at the same time, and will not produce accumulation, so it
588 should not be 30 years. (哪吒是一个个体，假设哪吒的三个头都得了脑血栓，那应该是同时发
589 生的事情，并不会产生累加，所以不应该是三十年。)

590 C.3 Definition and Examples of Score 3

591 **Definition** The LLM’s response correctly points out the type of irrationality/humor, but the key
592 points of irrationality/humor pointed out by the correct answer are inconsistent. Alternatively, the
593 response is consistent with the key points of irrationality/humor pointed out by the correct answer,
594 but the judgment on the type of cunning text is inconsistent.

595 C.3.1 The First Example of Score 3

596 **Input Text:** The drug is partially poisonous with side effects, so we should not take the drug in
597 parts. (是药三分毒，所以药不能分三次喝。)

598 **Model Response:** The humor of this sentence lies in the misunderstanding of the literal and actual
599 meaning of the idiom “The drug is partially poisonous with side effects”. This idiom is meant to say
600 that any drug has potential side effects, while the sentence misunderstands “partially” to mean that
601 the drug should not be taken in parts, and thus draws the absurd conclusion that the drug cannot be

602 taken in three separate doses, which is obviously a comical distortion of the original meaning. (这句
603 子的幽默之处在于对成语“是药三分毒”的字面意思和实际含义的误解。成语本意是说任何
604 药物都有潜在的副作用，而句子却把“三分”误解为药物应该分成三个部分服用，因此得出
605 了一个荒谬的结论，即药物不能按照三次分开服用，这显然是对原意的一种滑稽曲解。)

606 **Correct Answer:** “The drug is partially poisonous with side effects” refers to any drug will have
607 certain side effects if used improperly. “Partially” indicates that to a certain extent, there is no
608 meaning of “in parts”. Therefore, the conclusion “we should not take the drug in parts” cannot be
609 drawn from this sentence. (“是药三分毒”指的是任何的药品如果使用不当的话都会存在一定的
610 副作用，“三分”表示的是在一定程度上，并没有分三次的意思，所以“是药三分毒”这句话
611 并不能得出“药不能分三次喝”这一结论。)

612 C.3.2 The Second Example of Score 3

613 **Input Text:** Did a knee-jerk reflex test on a friend, and the result was that his leg never lifted again
614 after the hammer was knocked down. This reflex arc is too long! (给朋友做膝跳测试，结果锤下
615 去之后他的腿再也没抬起过，这反射弧也太长了吧！)

616 **Model Response:** The knee-jerk reflex test is a simple test to check the function of the nervous
617 system. Normally, a light tap on the tendon below the knee will cause the leg to make a quick leg
618 lift motion. This sentence humorously exaggerates the amount of time the friend does not respond,
619 indicating that the reflex arc is unusually long, when in fact, if there is really no response, it may
620 indicate a neurological problem rather than the reflex arc length. In addition, using “never lifted
621 again” has a more humorous effect because it indicates a permanent state, which is impossible in
622 reality. (膝跳反射测试是一种检查神经系统功能的简单测试，正常情况，轻轻敲击膝盖下方
623 的腱会导致腿部做出快速的抬腿动作。这句话幽默地夸张了朋友没有反应的时间，暗示反
624 射弧异常地长，实际上如果真的没有反应，那可能表明存在神经系统问题，而不是反射弧
625 长。此外，用“再也没抬起过”增加了幽默效果，因为它暗示了一种永久性的状态，这在现
626 实中是不可能的。)

627 **Correct Answer:** “His leg never lifted again after the hammer” indicates that the friend’s leg was
628 broken, and is not related to the length of the reflex arc, here is actually a humorous way to express
629 the matter of “breaking the friend’s leg”. (“锤下去之后他的腿再也没抬起过”说明把朋友的腿打
630 断了，与反射弧长短没有关系，这里实际上是用幽默的方式将“把朋友腿打断”表达出来。)

631 C.4 Definition and Examples of Score 4

632 **Definition** The response correctly points out the type of irrationality/humor in the input text, and
633 the key points pointed out are correct, but the interpretation is not consistent with the correct answer.

634 C.4.1 The First Example of Score 4

635 **Input Text:** In a vegetable market, one finds vegetative persons instead. So it must have been a job
636 market before. (去素材库发现几个植物人躺在里面，原来之前这里是人材市场。)

637 **Model Response:** The humor of this sentence lies in the pun on the word “vegetative”. Usually,
638 “vegetative” refers to patients who have fallen into a persistent vegetative state due to a major illness
639 or injury, but here it is used to refer to “vegetable” in the literal sense, i.e., real plants. At the same
640 time, “job market” usually refers to a market for hiring “people” with the appropriate skill, but here it
641 is used teasingly to refer to a market with vegetative “people”. This linguistic coincidence creates an
642 absurd humorous effect. (这个句子的幽默之处在于使用了“植物人”这个词语的双关意味。通
643 常，“植物人”指的是那些因为重大疾病或伤害而陷入持续植物状态的病人，但在这里它被
644 用来指代字面意义上的“植物”，即真正的植物。同时，“人材市场”通常是指招聘人才的市场，
645 而在这里它被戏谑地用来指一个植物市场，这种语言上的巧合创造了一种荒谬的幽默
646 效果。)

647 **Correct Answer:** The vegetable market refers to the market where vegetables are sold. “Vegetative”
648 is a special human state similar to the living state of plants, and the human body cannot lie in the
649 vegetable market. (素材库是指销售蔬菜的市场。“植物人”是与植物生存状态相似的特殊的人
650 体状态，人体是无法躺进素材库中的。)

651 C.4.2 The Second Example of Score 4

652 **Input Text:** If you drink 82 Lafite, then you’ll live at least 82 years. (喝了82年拉菲至少能
653 活82岁。)

654 **Model Response:** The humor or irrationality of the sentence is that it establishes a false cause-and-
655 effect relationship, implying that drinking a certain vintage of wine guarantees living to the same
656 age as the vintage, which obviously has no scientific basis. In fact, a person’s lifespan is affected by
657 many factors, including genetics, lifestyle, medical conditions, and so on, rather than a single diet.
658 In addition, this statement also carries a kind of humorous exaggeration, because it uses an absurd
659 logic to attract attention and increase the interest of the words. (该句子的幽默或不合理之处在于
660 它建立了一种错误的因果关系，暗示喝了某年份的葡萄酒能保证活到与年份相同的年龄，
661 这显然没有科学依据。实际上，人的寿命受多种因素影响，包括遗传、生活方式、医疗条
662 件等，而非单一的饮食习惯。此外，这种说法也带有一种幽默的夸张，因为它用一种荒谬
663 的逻辑来吸引注意，增加话语的趣味性。)

664 **Correct Answer:** “82 Lafite” is a network buzzword that originated from scenes of opening Lafite
665 wine in movies and TV series. Since the Lafite red wine produced in 1982 is of good quality and
666 expensive, “82 Lafite” is synonymous with high standards to set off the character’s extravagant
667 spending and taste, rather than “Lafite has been drunk for 82 years”. (82年的拉菲，网络流行
668 语，源自影视剧中开拉菲的桥段，由于82年的拉菲红酒品质好、价格昂贵，所以“82年的拉
669 菲”是高规格的代名词，来衬托人物的挥金如土和品味，而不是“喝了八十二年的拉菲”的意
670 思。)

671 C.5 Definition and Examples of Score 5

672 **Definition** The LLM’s response and the correct explanation are basically matched, the response
673 correctly points out the type of unreasonable/humorous points in the input text, the key points are
674 correct, and the explanation is basically consistent with the explanation.

675 C.5.1 The First Example of Score 5

676 **Input Text:** If a police officer commits a crime and turns himself in, is it considered to have made
677 meritorious contributions? (警察犯了事去自首算不算戴罪立功啊?)

678 **Model Response:** “Meritorious service” usually refers to the behavior of criminals who actively
679 cooperate with the police to detect other cases or fight criminal activities after committing a crime, so
680 as to reduce their own punishment. If the police officer has committed a crime and turned himself in,
681 and there is no further meritorious service, then this can usually only be counted as surrender, and
682 cannot be directly referred to as meritorious service. However, surrendering may lead to reduced
683 punishment according to relevant legal provisions. (“戴罪立功”通常指的是犯罪分子在犯罪后，
684 积极配合警方侦破其他案件或对抗犯罪活动，从而争取减轻自己刑罚的行为。如果警察犯
685 了罪后去自首，并且没有进一步的立功表现，那么这通常只能算作自首，并不能直接称为
686 戴罪立功。不过，自首可能会根据相关法律规定减轻处罚。)

687 **Correct Answer:** According to the provisions of Article 67 of the Criminal Law, “Surrender”
688 refers to the act of automatically surrendering to the police after committing a crime and truthfully
689 confessing one’s crime. If a criminal suspect or defendant under compulsory measures or a criminal
690 serving a sentence truthfully confesses other crimes that are not yet known to the judicial organs,
691 he shall be regarded as surrendering. “Meritorious service” means that the meritorious service

692 performance in criminal law refers to the performance of criminal suspects providing important
693 clues, exposing others' crimes, etc., and can be mitigated after verification. Surrender confession by
694 oneself is one's own crime, while meritorious service is to expose the crime of others, so the police
695 committing the crime to surrender is not regarded as meritorious service. (根据刑法第67条的规
696 定,“自首”是指犯罪以后自动投案,如实供述自己的罪行的行为。被采取强制措施的犯罪
697 嫌疑人、被告人和正在服刑的罪犯,如实供述司法机关还未掌握的本人其他罪行的,以自
698 首论。“戴罪立功”就是指刑法中的立功表现是指犯罪嫌疑人提供重要线索、揭发他人犯罪
699 行等的表现,经查实的可以获得减轻处罚。自首供述的是本人的罪行,戴罪立功是揭发他
700 人罪行,所以条子犯了事去自首不算戴罪立功。)

701 C.5.2 The Second Example of Score 5

702 **Input Text:** Officer Wang shot the hostage and saved the robber who could have been killed at any
703 time. (王警官击毙人质,解救了随时可能被狙杀的劫匪。)

704 **Model Response:** The unreasonableness of this sentence lies in the fact that Officer Wang's actions
705 are contrary to the duties of ordinary police officers. Usually, the task of the police is to protect the
706 hostages and try to arrest the criminals, not to kill the hostages to protect the criminals. Thus, the
707 sentence reverses the duties of the police with an absurd logic that produces a humorous effect. (这
708 句子的不合理之处在于王警官的行动与一般警察的职责相反。通常情况下,警察的任务是
709 保护人质并尽量逮捕犯罪分子,而不是击毙人质来保护犯罪分子。因此,这个句子以一种
710 荒谬的逻辑颠倒了警察的职责,产生了幽默效果。)

711 **Correct Answer:** As a police officer, what Wang should do is shoot the robber to save the hostage,
712 not shoot the hostage to save the robber. The fact that Wang killed the hostage by mistake is described
713 humorously. (王警官作为警察,应该做的是击毙劫匪解救人质,而不是击毙人质解救劫匪,
714 这里用幽默的方式阐述了王警官误杀人质这一事实。)

715 D Datasheet For FLUB

716 D.1 Motivation

717 1. **For what purpose was the dataset created?** Was there a specific task in mind? Was there
718 a specific gap that needed to be filled? Please provide a description.

719 • In this work, we aim to challenge the reasoning and understanding abilities of LLMs
720 by proposing the FLUB containing cunning texts that are easy for humans to understand
721 but difficult for models to grasp. Specifically, we design three tasks with increasing
722 difficulty to test whether the LLMs can understand the fallacy and solve the “cunning”
723 texts: Answer Selection, (2) Cunning Type Classification, (3) Fallacy Explanation. We
724 hope and believe that our proposed FLUB and all our findings are crucial for LLMs to
725 comprehend the fallacy and handle cunning texts in the real world.

726 2. **Who created the dataset (e.g., which team, research group) and on behalf of which**
727 **entity (e.g., company, institution, organization)?**

728 • The dataset is presented by Tsinghua Knowledge Engineering Laboratory (SZ).

729 3. **Who funded the creation of the dataset?** If there is an associated grant, please provide the
730 name of the grantor and the grant name and number.

731 • This work is sponsored by NSFC, Guangdong Province, Shenzhen City, Peng Cheng
732 Laboratory, and Tsinghua University.

733 4. **Any other comments?**

734 • No.

735 **D.2 Composition**

- 736 **5. What do the instances that comprise the dataset represent (e.g., documents, photos,**
737 **people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings;
738 people and interactions between them; nodes and edges)? Please provide a description.
- 739 • All the instances in FLUB are represented by texts. We make our benchmark openly
740 available on the GitHub page (<https://github.com/THUKElab/FLUB>).
- 741 **6. How many instances are there in total (of each type, if appropriate)?**
- 742 • FLUB includes 834 instances. For fine-grained cunning types, “False Analogy” has 11
743 instances, “Lame Jokes” has 44 instances, “Phonetic Error” has 5 instances, “Ambi-
744 guity” has 35 instances, “Paradox” has 29 instances, “Factual Error” has 12 instances,
745 “Reasoning Error” has 445 instances, “Word Game” has 239 instances, and “Undefined”
746 has 14 instances.
- 747 **7. Does the dataset contain all possible instances or is it a sample (not necessarily random)**
748 **of instances from a larger set?** If the dataset is a sample, then what is the larger set? Is the
749 sample representative of the larger set (e.g., geographic coverage)? If so, please describe
750 how this representativeness was validated/verified. If it is not representative of the larger set,
751 please describe why not (e.g., to cover a more diverse range of instances, because instances
752 were withheld or unavailable).
- 753 • FLUB has contained all possible instances, because we have tried our best to collect as
754 much data as possible from “Ruozhiba” and conducted strict manual annotation.
- 755 **8. What data does each instance consist of? “Raw” data (e.g., unprocessed text or images)**
756 **or features?** In either case, please provide a description.
- 757 • Each instance consists of the input cunning text, cunning type, fallacy explanation,
758 candidate answers, and the corresponding correct option, as illustrated in Figure 1b.
- 759 **9. Is there a label or target associated with each instance?** If so, please provide a description.
- 760 • There is a cunning type for each instance, which describes the cunning type of each
761 input text.
- 762 **10. Is any information missing from individual instances?** If so, please provide a description,
763 explaining why this information is missing (e.g., because it was unavailable). This does not
764 include intentionally removed information, but might include, e.g., redacted text.
- 765 • No.
- 766 **11. Are relationships between individual instances made explicit (e.g., users’ movie ratings,**
767 **social network links)?** If so, please describe how these relationships are made explicit.
- 768 • Not applicable.
- 769 **12. Are there recommended data splits (e.g., training, development/validation, testing)?** If
770 so, please provide a description of these splits, explaining the rationale behind them.
- 771 • No, because FLUB is a benchmark test set, all its instances are used for testing LLMs,
772 regardless of training/validation.
- 773 **13. Are there any errors, sources of noise, or redundancies in the dataset?** If so, please
774 provide a description.
- 775 • No.
- 776 **14. Is the dataset self-contained, or does it link to or otherwise rely on external resources**
777 **(e.g., websites, tweets, other datasets)?** If it links to or relies on external resources, a) are
778 there guarantees that they will exist, and remain constant, over time; b) are there official
779 archival versions of the complete dataset (i.e., including the external resources as they
780 existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees)
781 associated with any of the external resources that might apply to a dataset consumer? Please
782 provide descriptions of all external resources and any restrictions associated with them, as
783 well as links or other access points, as appropriate.

- 784 • FLUB is self-contained.
- 785 15. **Does the dataset contain data that might be considered confidential (e.g., data that is**
786 **protected by legal privilege or by doctor–patient confidentiality, data that includes the**
787 **content of individuals’ non-public communications)?** If so, please provide a description.
- 788 • No.
- 789 16. **Does the dataset contain data that, if viewed directly, might be offensive, insulting,**
790 **threatening, or might otherwise cause anxiety?** If so, please describe why.
- 791 • No. We have conducted a strict data cleaning process to ensure that FLUB does not
792 contain unethical data.
- 793 17. **Does the dataset identify any subpopulations (e.g., by age, gender)?** If so, please
794 describe how these subpopulations are identified and provide a description of their respective
795 distributions within the dataset.
- 796 • No.
- 797 18. **Is it possible to identify individuals (i.e., one or more natural persons), either directly or**
798 **indirectly (i.e., in combination with other data) from the dataset?** If so, please describe
799 how.
- 800 • No.
- 801 19. **Does the dataset contain data that might be considered sensitive in any way (e.g.,**
802 **data that reveals race or ethnic origins, sexual orientations, religious beliefs, political**
803 **opinions or union memberships, or locations; financial or health data; biometric or**
804 **genetic data; forms of government identification, such as social security numbers;**
805 **criminal history)?** If so, please provide a description.
- 806 • No.
- 807 20. **Any other comments?**
- 808 • No.

809 D.3 Collection Process

- 810 21. **How was the data associated with each instance acquired? Was the data directly observ-**
811 **able (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indi-**
812 **rectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses**
813 **for age or language)?** If the data was reported by subjects or indirectly inferred/derived
814 from other data, was the data validated/verified? If so, please describe how.
- 815 • We collect raw text data from “Ruozhiba” in Baidu Tieba, as described in Section 2.1.
816 The data is directly observable at <https://github.com/THUKElab/FLUB>.
- 817 22. **What mechanisms or procedures were used to collect the data (e.g., hardware appara-**
818 **tuses or sensors, manual human curation, software programs, software APIs)? How**
819 **were these mechanisms or procedures validated?**
- 820 • We use web crawlers to automatically crawl the raw data, and we perform manual
821 filtering and filtering to validate the crawled data, as described in Section 2.1.
- 822 23. **If the dataset is a sample from a larger set, what was the sampling strategy (e.g.,**
823 **deterministic, probabilistic with specific sampling probabilities)?**
- 824 • Not applicable.
- 825 24. **Who was involved in the data collection process (e.g., students, crowdworkers, contrac-**
826 **tors) and how were they compensated (e.g., how much were crowdworkers paid)?**
- 827 • We hired crowdworkers to clean the raw data and paid each person \$0.50 per piece of
828 raw data.

- 829 25. **Over what timeframe was the data collected?** Does this timeframe match the creation
830 timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?
831 If not, please describe the timeframe in which the data associated with the instances was
832 created.
- 833 • The raw data of FLUB was collected in in October 2023. The task characteristics of FLUB
834 are not time-sensitive, so the collection time is not associated with the data instances.
- 835 26. **Were any ethical review processes conducted (e.g., by an institutional review board)?**
836 If so, please provide a description of these review processes, including the outcomes, as well
837 as a link or other access point to any supporting documentation.
- 838 • Not applicable. Our data collection process does not involve human or animal experi-
839 ments. In addition, according to the Baidu Bar agreement, the data on Baidu Tieba can
840 be used for academic research free of charge and without liability. Therefore, our data
841 collection process does not require the involvement of an ethical review board.
- 842 27. **Did you collect the data from the individuals in question directly, or obtain it via third**
843 **parties or other sources (e.g., websites)?**
- 844 • We collect raw text data from “Ruozhiba” in Baidu Tieba, as described in Section 2.1.
- 845 28. **Were the individuals in question notified about the data collection?** If so, please describe
846 (or show with screenshots or other information) how notice was provided, and provide a link
847 or other access point to, or otherwise reproduce, the exact language of the notification itself.
- 848 • Not applicable. According to the Baidu Bar agreement, the data on Baidu Tieba can be
849 used for academic research free of charge and without liability.
- 850 29. **Did the individuals in question consent to the collection and use of their data?** If so,
851 please describe (or show with screenshots or other information) how consent was requested
852 and provided, and provide a link or other access point to, or otherwise reproduce, the exact
853 language to which the individuals consented.
- 854 • Yes. According to the Baidu Bar agreement, the data on Baidu Tieba can be used for
855 academic research free of charge and without liability.
- 856 30. **If consent was obtained, were the consenting individuals provided with a mechanism to**
857 **revoke their consent in the future or for certain uses?** If so, please provide a description,
858 as well as a link or other access point to the mechanism (if appropriate).
- 859 • Not applicable.
- 860 31. **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g.,**
861 **a data protection impact analysis) been conducted?** If so, please provide a description of
862 this analysis, including the outcomes, as well as a link or other access point to any supporting
863 documentation.
- 864 • Yes. The cunning texts we are concerned about come from daily life and are very
865 common. Therefore, the new research direction and tasks we propose will not cause
866 harm to human society.
- 867 32. **Any other comments?**
- 868 • No.

869 D.4 Preprocessing/cleaning/labeling

- 870 33. **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucket-**
871 **ing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances,**
872 **processing of missing values)?** If so, please provide a description. If not, you may skip the
873 remaining questions in this section.
- 874 • Yes. We employ annotators to manually filter out irrelevant posts that do not present
875 cunning texts. Since the collected original posts contain irrelevant content such as links

876 and images, we also require annotators to extract the fallacious and illogical contents
877 from the raw post and rewrite them into a complete sentence. Besides, it is worth
878 noting that we carefully ensure that the texts in FLUB are ethical texts. This process
879 includes user information anonymization, sensitive information removal, and filtering
880 of impolite posts.

881 34. **Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to**
882 **support unanticipated future uses)?** If so, please provide a link or other access point to
883 the “raw” data.

884 • No.

885 35. **Is the software that was used to preprocess/clean/label the data available?** If so, please
886 provide a link or other access point.

887 • No.

888 36. **Any other comments?**

889 • No.

890 D.5 Uses

891 37. **Has the dataset been used for any tasks already?** If so, please provide a description.

892 • No.

893 38. **Is there a repository that links to any or all papers or systems that use the dataset?** If
894 so, please provide a link or other access point.

895 • No.

896 39. **What (other) tasks could the dataset be used for?**

897 • Based on our constructed FLUB and its annotation information, we design three tasks
898 with increasing difficulty to test whether the LLMs can understand the fallacy and
899 solve the “cunning” texts. Specifically, (1) Answer Selection: The model is asked to
900 select the correct one from the four answers provided by FLUB for each input text. (2)
901 Cunning Type Classification: Given a cunning text as input, the model is expected to
902 directly identify its fallacy type defined in our scheme. (3) Fallacy Explanation: We
903 hope the model sees a cunning text and intelligently generates a correct explanation for
904 the fallacy contained in the text, just like humans, without falling into its trap.

905 40. **Is there anything about the composition of the dataset or the way it was collected**
906 **and preprocessed/cleaned/labeled that might impact future uses?** For example, is there
907 anything that a dataset consumer might need to know to avoid uses that could result in unfair
908 treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks
909 or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there
910 anything a dataset consumer could do to mitigate these risks or harms?

911 • No.

912 41. **Are there tasks for which the dataset should not be used?** If so, please provide a
913 description.

914 • According to the characteristics of the data in FLUB, it is known that in addition to the
915 three benchmark tasks we designed, we think that it may also be suitable for improving
916 the reasoning ability and humor ability of LLMs. Beyond that, FLUB may not be
917 suitable for other tasks.

918 42. **Any other comments?**

919 • No.

920 D.6 Distribution

- 921 43. **Will the dataset be distributed to third parties outside of the entity (e.g., company,**
922 **institution, organization) on behalf of which the dataset was created?** If so, please
923 provide a description.
- 924 • Yes, the dataset has been open-source.
- 925 44. **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** Does
926 the dataset have a digital object identifier (DOI)?
- 927 • The data is available through <https://github.com/THUKElab/FLUB>.
- 928 45. **When will the dataset be distributed?**
- 929 • The dataset has been open-source.
- 930 46. **Will the dataset be distributed under a copyright or other intellectual property (IP)**
931 **license, and/or under applicable terms of use (ToU)?** If so, please describe this license
932 and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant
933 licensing terms or ToU, as well as any fees associated with these restrictions.
- 934 • FLUB is published under [Creative Commons Attribution 4.0 International \(CC BY 4.0\)](#),
935 which means everyone can use this dataset for non-commercial research purposes.
- 936 47. **Have any third parties imposed IP-based or other restrictions on the data associated**
937 **with the instances?** If so, please describe these restrictions, and provide a link or other
938 access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees
939 associated with these restrictions.
- 940 • We collect raw text data from “Ruozhiba” in [Baidu Tieba](#). According to the [Baidu Bar](#)
941 [agreement](#), the data on Baidu Tieba can be used for academic research free of charge
942 and without liability.
- 943 48. **Do any export controls or other regulatory restrictions apply to the dataset or to**
944 **individual instances?** If so, please describe these restrictions, and provide a link or other
945 access point to, or otherwise reproduce, any supporting documentation.
- 946 • No.
- 947 49. **Any other comments?**
- 948 • No.

949 D.7 Maintenance

- 950 50. **Who will be supporting/hosting/maintaining the dataset?**
- 951 • Tsinghua Knowledge Engineering Laboratory (SZ) will support hosting of the dataset.
- 952 51. **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
- 953 • The manager of FLUB can be contacted through:
 - 954 • Email (liyinghu20@mails.tsinghua.edu.cn)
 - 955 • GitHub issues (<https://github.com/THUKElab/FLUB/issues>).
- 956 52. **Is there an erratum?** If so, please provide a link or other access point.
- 957 • There is no erratum for our first release. Errata will be documented on the dataset
958 website as a future release.
- 959 53. **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete in-**
960 **stances)?** If so, please describe how often, by whom, and how updates will be communicated
961 to dataset consumers (e.g., mailing list, GitHub)?
- 962 • Yes. Once any other researchers find that FLUB needs to be updated, we will immediately
963 update it through GitHub.

964 54. **If the dataset relates to people, are there applicable limits on the retention of the data**
965 **associated with the instances (e.g., were the individuals in question told that their data**
966 **would be retained for a fixed period of time and then deleted)?** If so, please describe
967 these limits and explain how they will be enforced.

- 968 • Not applicable.

969 55. **Will older versions of the dataset continue to be supported/hosted/maintained?** If so,
970 please describe how. If not, please describe how its obsolescence will be communicated to
971 dataset consumers.

- 972 • Yes. We will continue to support FLUB. Once any other researchers find that FLUB needs
973 to be updated, we will immediately update it through GitHub.

974 56. **If others want to extend/augment/build on/contribute to the dataset, is there a mecha-**
975 **nism for them to do so? If so, please provide a description. Will these contributions**
976 **be validated/verified?** If so, please describe how. If not, why not? Is there a process for
977 communicating/distributing these contributions to dataset consumers? If so, please provide
978 a description.

- 979 • Yes. Once any other researchers find that FLUB needs to be updated, after they contact
980 us via email or GitHub, we will review the data they want to expand. After the new
981 data passes review, we will immediately update it to GitHub.

982 57. **Any other comments?**

- 983 • No.

984 E Metadata and Data Format of FLUB

985 E.1 Croissant Metadata

986 To provide the key descriptive information of FLUB more clearly and improve the traceability
987 and reproducibility of our data, we also provide the Croissant metadata of FLUB, please refer
988 to the link https://github.com/THUKElab/FLUB/blob/main/FLUB_croissant_metadata.json
989 for details.

990 E.2 Data Format

Listing 1: The data format of our FLUB.

```
991 { 1
992   2
993   "text": "The input cunning text", 3
994   "is_question": "Is the input cunning text a question?", 4
995   "type": "The cunning type of the input text for the Cunning Type 5
996     Classification task.", 6
997   "explanation": "The correct explanation of the input text for 7
998     the Fallacy Explanation task.", 8
999   "id": "The id of each data sample", 9
1000   "options": { 10
1001     "A": "The candidate answer 1 for the input text (question)", 11
1002     "B": "The candidate answer 2 for the input text (question)", 12
1003     "C": "The candidate answer 3 for the input text (question)", 13
1004     "D": "The candidate answer 4 for the input text (question)" 14
1005   },
1006   "answer": "The correct answer for the Answer Selection (Multiple 13
1007     Choice) task." 14
1008 }
```

1010 **F Author Statement of FLUB**

1011 We, as the authors of the FLUB dataset, hereby declare the following:

- 1012 1. **Responsibility Statement:** The creation, organization, and publication of FLUB are entirely
1013 our responsibility. We confirm that all data were legally obtained and do not infringe on
1014 the intellectual property or other rights of any third party. In the event of any disputes or
1015 legal liabilities arising from the use of this dataset, we, as the authors, will assume full
1016 responsibility.
- 1017 2. **Data License:** This dataset is released under the following license: Creative Commons
1018 Attribution 4.0 International (CC BY 4.0). Users must comply with the terms of this license
1019 agreement when using this dataset. For detailed license terms, please refer to [CC BY 4.0](#).
- 1020 3. **Data Integrity and Quality:** We have made every effort to ensure the integrity and quality
1021 of FLUB. However, due to the dataset's size and complexity, we cannot guarantee it to be
1022 completely error-free. If any errors or omissions are discovered, please contact us for
1023 corrections and updates.
- 1024 4. **Ethical Statement:** We have strictly adhered to relevant ethical guidelines during the data
1025 collection and processing stages to ensure that the use of this dataset does not negatively
1026 impact or harm any individual or organization.