NAIL: A CHALLENGING BENCHMARK FOR NAÏVE LOGICAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Logical reasoning over natural text is an important capability towards human level intelligence. Existing datasets are either limited and inadequate to train and evaluate logical reasoning capability (e.g., LogiQA and ReClor), or not oriented for logical reasoning (e.g., SQuAD and HotpotQA). In this paper, we focus on a specific category of logical reasoning, named *naïve logical reasoning*, and propose a new large scale benchmark, named NAIL, targeted for learning and evaluating models' capabilities towards naïve logical reasoning. NAIL is source from standardized exams such as Chinese National Civil Servants Examination and Law School Admission Test. Furthermore, to collect more data, we propose to imitate the example of standardized exams rather than designing them from scratch. NAIL is available in both Chinese and English containing a total of $10,296 * 2$ instances. Empirical results show that current state-of-the-art neural models struggle on NAIL with very poor accuracy (the best result is 30.10% for NAIL and 36.15% for Chinese NAIL), while human experts can perform nearly 100% accuracy. Further results indicate that human imitations can significantly help models learn logic from natural text.

1 INTRODUCTION

Current deep models achieve near human-level performance on many tasks in NLP (Devlin et al., 2019; Liu et al., 2019), but often surface-level knowledge suffices to answer. To move towards human intelligence, we need to equip the models with logical reasoning capabilities (e.g., ability to draw logical conclusions from given statements), which is also a long sought-after goal of AI (Newell & Simon, 1956; McCarthy et al., 1960). One related task is *natural language inference* whose goal is to assign the logical relationships (`contradicted`, `neutral` and `entailment`) of sentence pairs. To push the development of models in logical reasoning, the researchers have focused on more challenging reading comprehension tasks, which often require more complex reasoning as well as longer input. However, most existing reading comprehension datasets are not oriented for the logical reasoning (e.g., SQuAD and HotpotQA), with the exception of LogiQA (Liu et al., 2020), ReClor (Yu et al., 2020) and LR-LSAT (Wang et al., 2021).

Above datasets (LogiQA, ReClor and LR-LSAT) are limited and inadequate to train and evaluate logical reasoning capability. The reason is that all of the three datasets involve diverse types of logical reasoning, such as drawing an alternate conclusion to the argument, finding necessary/sufficient assumptions, whether statements strengthen/weaken the argument or explain/resolve the situation. Mixing multiple types of logical reasoning may pose the following challenges. a). From the perspective of human cognition, different types of logical reasoning correspond to different problem-solving ideas. But in practice we usually train a model on the whole dataset with the same idea, which makes it more limited for models to learn logical reasoning capability. b). From the perspective of machine learning, if there are many reasoning types mixed in a dataset, then there will be less data for each reasoning type, which is inadequate to train and evaluate logical reasoning capability (demonstrated in our experiments). Furthermore, when the model does not work, it is difficult to determine which reasoning type is the bottleneck (no reasoning type annotation in the dataset), which may hinder the design of better models.

To tackle the challenges, we focus on a more fine-grained type of logical reasoning, naïve logical reasoning (Section 2), which is to infer the logical conclusion from statements that describe triples

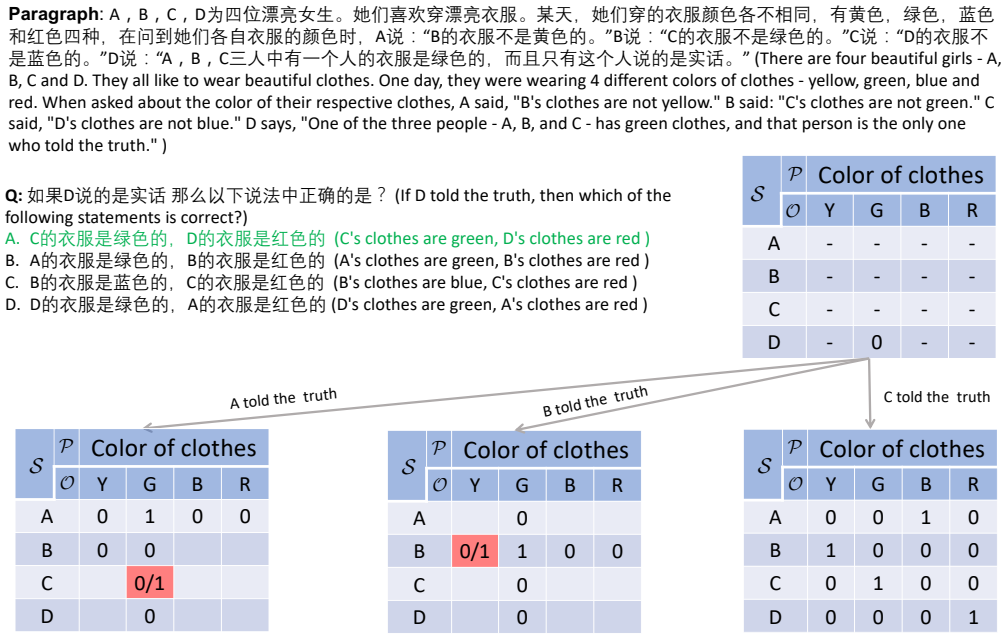

Figure 1: An example of NAIL requiring naïve logical reasoning. Green font indicates the correct answer. Tables depict the reasoning process from the human perspective. - indicates uncertain state. 0 and 1 indicate true value. Red (0/1) means there is a conflict.

(subject, predicate, object) and the relationships among them. A typical example of naive logical reasoning is shown in Figure 1. To answer the query in this example, we need to derive them separately by condition, and stop the branch if a conflict occurs. Specifically, in the initial table, we only know that "D is not green" from "D told the truth". Assume that "A told the truth", and infer that "A is green, C is non-green". Next "C is green" from that "C told the lie". C derives both green and non-green, thus causing a conflict. Similar processes for "B told the truth" and "C told the truth". Human minds need extensive training and practice to get used to such complex logical reasoning.

Inspired by the datasets extracted from standardized examinations (Lai et al., 2017; Clark et al., 2018; Liu et al., 2020), we build a new large scale benchmark, NAIL, by selecting naive logical reasoning examples from standardized exams such as the National Civil Servants Examination of China and Law School Admission Test. However, examples of naive logical reasoning are limited due to the immense effort required by human experts in design. To collect more data, we propose to imitate the example of standardized exams rather than designing them from scratch. Unlike simple data augmentation (e.g., substitution, paraphrasing), human imitation aims at getting more diverse examples while maintaining the underlying logic (Figure 1). NAIL is available in both Chinese and English, containing a total of $10,296 * 2$ instances.

Empirical results show that current state-of-the-art neural models struggle on NAIL with very poor accuracy (the best result is 30.10% for NAIL and 36.15% for Chinese NAIL), while human experts can perform nearly 100% accuracy. Further results indicate that human imitations can significantly help models learn logic from natural text. However, the performance of models on transformation instances (e.g., phrase/statement-level shuffling and paraphrasing of sentences) is close to that of random guesses, suggesting that more exploration is needed to enhance current models to learn robust logic from diverse natural text.

2 NAÏVE LOGICAL REASONING

The naïve logical reasoning is a more fine-grained type of logical reasoning. Formally, we give the definition as follows.

Definition 1. *The subject is a described resource, usually an entity, such as person or location. The predicate indicates an attribute of the subject or indicates some kind of relationship between the subject and the object. When denoting an attribute, the object is the attribute value, usually a literal value, e.g., (Jacket, Color, Red), otherwise the object is an entity, e.g., (Beijing, Capital, China).*

Definition 2. *Assuming that the set of subjects $\mathcal{S}$, the set of predicates $\mathcal{P}$, the set of objects $\mathcal{O}$ and several statements, each statement describes a triple (subject, predicate, object) or some logical relationship between triples. The* ***naive logical reasoning*** *is the process of reasoning from these statements to reach a logical conclusion.* [1]

In this work, we explore naive logical reasoning in the form of reading comprehension. A typical example and detailed reasoning process is shown in Figure 1. Similar to the format of multiple-choice reading comprehension, it contains a context, a query and four options with only one correct answer. To solve the example, the model needs to understand the logical connections between the subjects, the predicates and the objects, then derive a valid option.

3 NAIL: DATA COLLECTION AND ANALYSIS

NAIL is a carefully designed benchmark for naïve logical reasoning similar to the format of multiple-choice reading comprehension. Inspired by the datasets extracted from standardized examinations (Lai et al., 2017; Clark et al., 2018; Liu et al., 2020), we first collect a small amount of examples from examinations, denoted NAIL-E (Section 3.1). Then we propose to imitate the examples of NAIL-Examination to collect more data, denoted NAIL-I (Section 3.2). Finally we provide a detailed analysis of the proposed NAIL (Section 3.3).

3.1 COLLECTION FROM EXAMINATION

We searched for such examples widely from two different types of public examinations: Chinese National Civil Servants Examination (CNCSE) and Law School Admission Test (LSAT). [2] CNCSE is a hotly contested annual event, there are overall 120-140 examples per exam per year. But only 1-4 examples belong to the scope of naïve logical reasoning, which is also the most difficult type of problems for candidates within the given 120 minutes. And the LSAT is a standardized test for prospective law school candidates in the United States, Canada, and a growing number of other countries. Logical Reasoning is a multiple-choice section of LSAT, containing 24-26 questions under the limitation of 35 minutes, where about 2-4 problems fall into naïve logical reasoning category.

We artificially selected examples that belong to the naïve logical reasoning category from the above two examinations. Finally we obtained 488 examples from the last 25 years of CNCSE and LSAT in the last 30 years, denoted NAIL-E. These two exams are from countries with different native languages. The former is expressed in Chinese, and the latter is expressed in English. And note that there are slight difference in language style between them. And for diversity and fairness, in later step we get all examples in both Chinese and English through translation.

3.2 COLLECTION FROM IMITATION

After collecting a small amount of NAIL-E, we expect to expand the amount of dataset. Designing examples from scratch requires a huge effort from human experts. One simple solution is data augmentation, which artificially scales up data by creating modified data from existing data, such as word/sentence shuffling, word replacement and syntactic variation. However, The data augmented in this way is highly correlated with the original data, and the model easily captures these semantic surface correlations, which makes it limited to train and evaluate logical reasoning capability. To alleviate this limitation, we propose to imitate the examples of NAIL-E, and create more examples with a diverse semantic surface while keeping the underlying logic of the original example. Furthermore, in the process of human imitation, we design strict strategies to control the quality of imitation.

Imitation Example See Figure 2, the context of example from NAIL-E consists of 6 sentences and a query (split by blank lines), each of which can be represented as a logical template (see "backbone template"). The example focus on the description of a scenario: `selecting Prince Charming` (m_1), which involves five entities: `Li_Na` (s_1), `Wang_Wei` (p_1), `Wu_Gang` (p_2), `Li_Qiang` (p_3),

[1] The term *naive* refers to the fact that this logical reasoning process of human in this task is spontaneous, intuitive and unsystematic, which is fundamentally based on the meta-theoretical concept of Naïve Set Theory Halmos (2017).

[2] `https://www.lsac.org/lsat/taking-lsat/test-format/logical-reasoning`

Backbone Template	Original Example	Imitation 1 ☹	Imitation 2 ☹	Imitation 3 ☺
s_1心中的 m_1 有如下特征：a_1, a_2, and a_3. (s_1's m_1 has following characters: a_1, a_2, and a_3.)	李娜心中的白马王子有如下特征：高个子、相貌英俊、博士。(Li Na's Prince Charming has following characters: tall, handsome, and a PhD.)	孙怡心中的白马王子有如下特征：高个子、相貌英俊、博士。(Sun Yi's Prince Charming has following characters: tall, handsome, and a PhD.)	孙怡心中的白马王子有如下特征：有钱、有房、有车。(Sun Yi's Prince Charming has following characters: rich, has a house, and has a car.)	小丽心中的理想礼物有如下特征：昂贵，有收藏价值，美观。(Xiao Li's ideal gift has following characters: expensive, valuable to collect, and nice-looking.)
她认识 p_1, p_2, p_3, p_4 4（量词）（p的类别）,其中有一符合s_1所要求的全部条件。(She knows 4 *type_of*(p_i), p_1, p_2, p_3, p_4, and one of them meets all the conditions s_1 requires.)	她认识王威、吴刚、李强、刘大伟4位男士,其中有一位符合她所要求的全部条件。(She knows 4 men, Wang Wei, Wu Gang, Li Qiang and Liu Dawei, and one of them meets all the conditions she requires.)	她认识瞿衡、梅震、季凡、张磊4位男士,其中有一位符合她所要求的全部条件。(She knows 4 men, Qu Heng, Mei Zhen, Ji Fan, and Zhang Lei, and one of them meets all the conditions she requires.)	她认识瞿衡、梅震、季凡、张磊4位男士,其中有一位符合她所要求的全部条件。(She knows 4 men, Qu Heng, Mei Zhen, Ji Fan, and Zhang Lei, and one of them meets all the conditions she requires.)	她知道手链、手表、手镯、戒指这4个礼物,其中有一个符合她所要求的全部条件。(She knows 4 gifts, bracelets, watches, bangles, and rings, and one of them meets all the conditions she requires.)
4（量词）（p的类别）中,有3（量词）是 a_1 ,有2（量词）是 a_3, 有1（量词）是a_2。(Among the 4 *type_of*(p_i), 3 are a_1, 2 are a_3, and 1 is a_2.)	(1)4位男士中,有3个高个子,2名博士,1人长相英俊。(Among the 4 men, 3 are tall, 2 are PhDs and 1 is handsome.)	(1) 4位男士中,有3个高个子,2名博士,1人长相英俊。(Among the 4 men, 3 are tall, 2 are PhDs and 1 is handsome.)	(1) 4位男士中,有3个有钱,2人有车,1人有房。(Among the 4 men, 3 are rich, 2 have a car and 1 has a house.)	(1) 4个礼物中,有3个很昂贵,2个很美观,1个具有收藏价值；(Among the 4 gifts, 3 are expensive, 2 are nice-looking and 1 is valuable to collect;)
p_1和 p_2都是 a_3。($p_1$and $p_2$are both a_3.)	(2)王威和吴刚都是博士。(Wang Wei and Wu Gang are both PhDs.)	(2) 瞿衡和梅震都是博士。(Qu Heng and Mei Zhen are both PhDs.)	(2) 瞿衡和梅震都有车。(Qu Heng and Mei Zhen both have a car.)	(2) 手链和手表都很美。(Bracelets and watches are both nice-looking.)
p_4和 p_3 关于a_1的属性相同。($p_4$and $p_3$are of the same *attribute_of*(a_1).)	(3)刘大伟和李强身高相同。(Liu Dawei and Li Qiang are of the same height.)	(3) 张磊和季凡身高相同。(Zhang Lei and Ji Fan are of the same height.)	(3) 张磊和季凡有同等程度的钱。(Zhang Lei and Ji Fan are of the same wealth.)	(3) 戒指和手镯价格相同。(Rings and bangles are of the same price.)
p_3和 p_1并非都是 a_1 。($p_3$and $p_1$are not both a_1 .)	(4)李强和王威并非都是高个子。(Li Qiang and Wang Wei are not both tall.)	(4) 季凡和瞿衡并非都是高个子。(Ji Fan and Qu Heng are not both tall.)	(4) 季凡和瞿衡并非都很富有。(Ji Fan and Qu Heng are not both rich.)	(4) 手镯和手链并非都是昂贵的。(Bangles and bracelets are not both expensive.)
请问谁符合s_1 要求的全部条件? (Who meets all the requirements of s_1 ?)	请问谁符合李娜要求的全部条件?(Who meets all the requirements of Li Na?)	请问谁符合孙怡要求的全部条件?(Who meets all the requirements of Sun Yi?)	请问谁符合孙怡要求的全部条件?(Who meets all the requirements of Sun Yi?)	请问哪个符合小丽要求的全部条件?(Which meets all the requirements of Xiao Li?)
A. p_4 *B.* p_3 *C.* p_2 *D.* p_1	A.刘大伟(Liu Dawei) B.李强(Li Qiang) C.吴刚(Wu Gang) D.王威(Wang Wei)	A.张磊(Zhang Lei) B.季凡(Ji Fan) C.梅震(Mei Zhen) D.瞿衡(Qu Heng)	A.张磊(Zhang Lei) B.季凡(Ji Fan) C.梅震(Mei Zhen) D.瞿衡(Qu Heng)	A.戒指(Rings) B.手镯(Bangles) C.手表(Watches) D.手链(Bracelets)

Figure 2: An original example from CNCSE and its three imitations, which share the same backbone template. Imitation 1 is Subject-level. Imitation 2 is Subject-and-Object-level. And Imitation 3 is Subject-and-Predicate-and-Object-level.

`Liu_Dawei` (p_4), and three noun/adjective properties: `tall` (a_1), `handsome` (a_2), `a_PhD` (a_3). An accepted imitation needs to keep original underlying logic but have semantically different groundings, i.e., to describe a completely different scenario. Imitation 1 and 2 are unqualified imitations. Since Imitation 1 only conducts Subject-level imitations (S imitations), i.e., only superficially substitute the five entities (s_1:`Li_Na`→`Sun_Yi`, p_1:`Wang_Wei`→`Qu_Heng`, p_2:`Wu_Gang`→`Mei_Zhen`, p_3:`Li_Qiang`→`Ji_Fan`, p_4:`Liu_Dawei`→`Zhang_Lei`). Further, Imitation 2 conducts Subject-and-Object-level imitations (SO imitations), i.e., not only substituting the entities, but also altering the corresponding objects (property values). (a_1:`tall`→`rich`, a_2:`handsome`→`having_a_house`, a_3:`a_PhD`→`having_a_car`). The difference between Imitation 2 and the original example is much greater than that between Imitation 1 and and the raw problem, however, Imitation 2 is still not an ideal imitation. Furthermore, Imitation 3 changes the scenario m_1 into `picking ideal gift`. Imitation 3 is an expected imitative writing, which conducts Subject-and-Predicate-and-Object-level imitations (SPO imitations). Note that all of the three imitations share the same logic templates. And p_2 is the answer for all of the original example and above three imitations.

Imitation Process We first select a group of people from a variety of occupations: professional editors, legal practitioners, in-service civil servants and college students (from different majors). Empirically, people in these occupations have strong logical and verbal skills. These people are asked to conduct SPO-level imitations based on the original example from NAIL-E, and reasonable imitations should meet two requirements: *logic invariance* and *semantic diversity*. We trained these candidate people how to conduct imitations such as Fig 2. We then conducted a trial phase before

the official imitation phase, in which process we eliminated some people of poor quality. Finally we employed 82 qualified people to imitatively construct problems based on NAIL-E, and they are paid RMB¥2.8 per imitation[3] Averagely, it costs a trained person about 3-4 minutes to finish an imitation–starting from coming up a scenario, then replacing the entities and attributes of the raw problem with those scenario-related while keeping invariant logic, and finally smoothing the new sentence with some transition words if necessary. For each original example in NAIL-E, we expect at least 20 imitations (except for extremely difficult cases). Overall we use 813 paid work hours in total to build the NAIL-I.

Quality Control We adopt following strategies to ensure the quality of imitations:

1. As mentioned above, we conducted a trial phase before the official imitation phase. In this phase, we asked them to imitate a small number of problems. Although we do not necessarily need them to write the backbone template, but we will check the logic and give feedback to help them understand the task. This process was iterated for three rounds. Only those who passed the trial phase could participate in the official imitation.

2. During the official imitation, we set up an online chat room to synchronize with employees and answer their questions timely.

3. To embrace *semantic diversity*, each original example is shown to at least 5 people, that is, one can only conduct 4 imitations based on one original example. People who are assigned to the same example imitate independently without interference from each other, to ensure varied inspiration for imitation.

4. To ensure *logic invariance*, we adopt a double-checking strategy:

   - Cross Checking: Everyday, for each employee, we sample 5 imitations from all of his/her daily imitations. And the sampled imitation is assigned to other 2 employees for cross-checking. The imitation will be qualified only if they both approved, and the criterion for approval is that the imitation share the same logics with the original example. If any one of the 5 imitations produced by one employee fail, then all imitations of that employee for that day will be returned to re-check and repair.
   - Post Checking: To further ensure that the underlying logic do not deviate during imitation, we collect another team of experts to solve the imitative examples. The team consists of 20 experienced experts, and each of them has been selected through the Chinese National Civil Servants Examination. Each imitative example was presented to 3 experts randomly, who are allowed to select one and only correct choice from "A", "B", "C", 'D', otherwise, "UNABLE_TO_ANSWER" if bugs exist in the example, causing no correct choice or multiple correct choices. As long as one of the 3 experts pointed out "UNABLE_TO_ANSWER", then the imitator of this problem should recheck the logic, until each of these 3 experts could give one and only correct choice from "A", "B", "C", 'D'. Note that in the post checking process, we broke up the imitations together and shuffle randomly, otherwise if a person is faced with imitations from same original example, he/she is prone to give a shortcut option with speculation.

Human Evaluation As mentioned above, an imitation is finally regarded as qualified only if the sampled 3 experts could all solve the example. For any example in NAIL-E, we also ask 3 experts in the team to solve it. Since the gold answer to a example in NAIL-E are provided in public by the examination committee, and corresponding imitations in NAIL-I share the same answer with the original example. Therefore, we calculated the mean accuracy of these three submissions on the overall NAIL, denoted as the performance of **human experts**. To better demonstrate the difficulty of NAIL, we also selected another team consisting of 20 first-year college students. Same as above, example is shown by 3 students randomly and they have to give the answer independently. We calculated the overall mean accuracy as the performance of **human baseline**. The human evaluation is paid separately. Each person can receive RMB¥1.5 for answering each example, which generally cost the person 2-3 minutes.

Since we hire part-time people to write imitatively rather than using the existing crowd-sourcing platform, we build our own website, where we can review the quality and overall progress at any time.

[3] A part-time employee can produce 15-20 imitations per hour, where he/she can get RMB¥42RMB¥56, while the local minimum wage is RMB¥23 per hour.

Paragraph + Question + Answers | **Reasoning Type**

Paragraph: 某省举行“文明城市”评比。4位评委对大家普遍看好的A、B、C等3城市获得“文明城市”称号的可能性进行了分析预测。评委甲说：“要么A市能获得，要么C市能获得。”评委乙说：“如果A市与C市能获得，则B市也能获得。”评委丙说：“只有当B市不能获得或者C市能获得时，A市才不能获得。”评委丁说：“我看B市能获得的可能性为零，而A市与C市一定能获得。”评比结束后发现，4位评委中只有一人预测成立。(In a competition for "civilized cities" held in a province, four judges predicted the likelihood that three cities(A, B and C) would be awarded the title of "civilized city". Judge A said, "Either City A or City C will win." Judge B said, "If City A and City C can get it, then City B can also get it." Judge C said, "City A can't get it only if City B can't get it or City C can get it." Judge D said, "I think the probability of City B getting it is zero, while City A and City C must get it." After the competition, only one of the four judges' predictions was found to be true.)
Q: 据此可以推出能获得“文明城市”称号的城市是? (What is the city that wins the title of "civilized city"?)
A. A市(City A)
B. B市(City B)
C. C市(City C)
D. A市，B市和C市(City A,B and C)

Real/Fake Mixture (36.03%)

Paragraph: 甲、乙和丙，一位是山东人，一位是河南人，一位是湖北人。现在只知道：丙比湖北人年龄大，甲和河南人不同岁，河南人比乙年龄小。(A, B and C, one is from Shandong, one is from Henan and one is from Hubei. Now we only know: C is older than the one from Hubei, A and the one from Henan are not the same age, and the one from Henan is younger than B.)
Q: 由此可以推知: (From this we can deduce:)
A. 甲不是湖北人(A is not from Hubei.)
B. 河南人比甲年龄小(The one from Henan is younger than A.)
C. 河南人比山东人年龄大(The one from Henan is older than that from Shandong.)
D. 湖北人年龄最小(The one from Hubei is the youngest.)

Ordering (20.40%)

Paragraph: 甲、乙、丙在北京、南京和成都工作，他们的职业是医生、演员和教师。已知：甲不在北京工作，乙不在南京工作；在北京工作的不是教师；在南京工作的是医生；乙不是演员。(A, B and C work in Beijing, Nanjing and Chengdu, and their occupations are doctor, actor and teacher. It's known that: A does not work in Beijing, B does not work in Nanjing; The one who works in Beijing is not a teacher; The one who works in Nanjing is a doctor; B is not an actor.)
Q: 那么,甲,乙,丙分别在哪里工作? (So, where do A, B and C work respectively?)
A. 南京,成都和北京(Nanjing, Chengdu and Beijing)
B. 成都,北京和南京(Chengdu, Beijing and Nanjing)
C. 南京,北京和成都(Nanjing, Beijing and Chengdu)
D. 成都,南京和北京(Chengdu, Nanjing and Beijing)

Matching (29.26%)

Paragraph: S市一所小学的学生户籍情况比较复杂，所有三年级学生的户籍都在本市，有些二年级学生的户籍也在本市，有些一年级学生是农民工子弟，而农民工子弟的户籍都不在本市。(The registered residence of students in an elementary school in city S is complicated. All students in grade 3 are registered in this city; some students in grade 2 are also registered in this city; some students in grade 1 are children of migrant workers, and none of the migrant workers' children are registered in this city.)
Q: 据此，可以推出：(According to this, it can be deduced that:)
A. 所有二年级学生都不是农民工子弟(all students in the grade 2 are not children of migrant workers)
B. 有些农民工子弟是三年级学生(some migrant workers' children are in grade 3)
C. 有些户籍在本市的学生是三年级学生(some students with registered residence in this city are in grade 3)
D. 有些一年级学生不是农民工子弟(some students in the grade 1 are not children of migrant workers)

Set Operation (14.31%)

Figure 3: The percentage and representative examples of each naive logical reasoning type.

	Train(NAIL)		**Dev**(NAIL)		**Test**(NAIL)	
	NAIL-E	NAIL-I	NAIL-E	NAIL-I	NAIL-E	NAIL-I
# Example	292	5906	97	1904	99	1998
# Ave./Max. of paragraph	68.7 / 155	70.2 / 188	67.3 / 151	70.3 / 212	67.0 / 149	73.3 / 185
# Ave./Max. of query	8.1 / 74	6.9 / 81	9.3 / 71	9.3 / 134	7.9 / 92	7.6 / 71
# Ave./Max. of option	9.2 / 74	9.7 / 130	9.1 / 65	8.5 / 134	8.6 / 76	7.7 / 97

Table 1: Data split and corresponding statistics.

3.3 DATA ANALYSIS

As mentioned above, we first collect NAIL-E from CNCSE and NAIL-I from LSAT, and then collect NAIL-I by human imitations. They are divided into training set, validation set and testing set with 6,198, 2,001 and 2097 data points respectively. The overall statistics of NAIL are summarized in Table 1. It is worth noting that the original example in NAIL-E and the corresponding imitations will only be in the same data split. We analyze and manually annotate the types of examples on the NAIL-E and group them into 4 categories including real/fake mixture, ordering, matching and set operation, whose percentages and representative examples are shown in Figure 3. Each of these types of examples requires naive logical reasoning.

4 EXPERIMENTS

4.1 BASELINE SYSTEMS

In this section, we describe the initial baselines that we evaluated on the NAIL dataset, including simple rule-based methods and strong neural-based methods using the pre-trained language models. We adopt the *accuracy* as the evaluation metric. Detailed hyper-parameters of the baseline models and the results on the validation set are listed in the Appendix.

Model	**Input**	NAIL		Chinese NAIL	
		Dev	**Test**	**Dev**	**Test**
Random	(C, Q, A)	25.0	25.0	25.0	25.0
Word Matching	(Q, A)	23.1	24.6	23.3	24.9
	(C, Q, A)	21.6	25.8	22.2	24.2
Sliding Window	(Q, A)	23.8	24.2	24.1	24.7
	(C, Q, A)	21.9	22.1	21.6	22.5
BERT $_{\text{LARGE}}$	(A)	24.55	24.73	27.23	29.07
	(Q, A)	29.87	29.11	30.18	27.95
	(C, Q, A)	28.99	29.38	29.28	27.71
RoBERTa $_{\text{LARGE}}$	(A)	27.32	26.45	29.36	27.65
	(Q, A)	27.81	27.85	32.63	30.47
	(C, Q, A)	34.63	30.10	37.28	36.15
Human baseline	(C, Q, A)	70.10	71.31	75.49	76.38
Human expert	(C, Q, A)	100	100	100	100

Table 2: Main results on NAIL (accuracy%). The column Input means whether to input context (C), query (Q) and answer options (A).

Rule-based Methods Following LogiQA (Liu et al., 2020), we adopt two simple rule-based methods based on text matching. Specifically, *word matching* is to measure the degree of unigram overlap between candidate answer and the given paragraph-question pair; *sliding window* takes into account the n-gram when calculating the matching score.

Neural-based Methods Pre-trained neural models, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), have achieved impressive performance on reading comprehension. The input to the pre-trained model is the concatenation of the paragraph, the question and the candidate answer, separated by `[SEP]` tokens, denoted as: `[CLS]`, paragraph, `[SEP]`, question, `[SEP]`, answer. After the encoding of the pre-trained model, the representation of `[CLS]` token followed by an multiple layer perceptron (MLP) is used for scoring. The all parameters are fine-tuned during training.

4.2 RESULTS AND ANALYSIS

The performance of all baselines on the NAIL is presented in Table 2. In particular, the human baseline is 71.31% and 76.38% on the test set of NAIL and Chinese NAIL separately, while the human expert is 100% since ambiguous examples are not included in the dataset. In comparison, all of the baselines perform much worse than humans, demonstrating that these methods are very limited in naïve logical reasoning. In addition, the results are relatively similar on the English and Chinese datasets. The rule-based approaches achieve an accuracy of 25.8% and 24.2%, which is similar to random guesses, indicating that word correlation can not be used to help improve performance. Pre-trained models have relatively good performance with about 4-10 points improvement compared to the rule-based approaches, showing that pre-trained models have a certain degree of commonsense and logical reasoning capabilities (Huang et al., 2019b). However, the best result by RoBERTa $_{\text{LARGE}}$ is 36.15% on the testing data of Chinese NAIL, which is still far below the human performance. This indicates that the knowledge in the pre-trained model is quite weak in logical reasoning.

We conduct experiments with different input settings. The setting of questions and answer options (Q, A) does not lead to significant improvements compared to the input setting of only answer options (A). One likely reason is that the queries usually do not provide much information, e.g., *According to this, it can be deduced that?* Further adding context gives rise to a noticeable boost, showing that the context is highly informative.

Trained on → **Evaluated on ↓**	NAIL	NAIL-E	NAIL-I
NAIL	30.10	26.39	33.93
NAIL-E	37.37	22.92	38.54
NAIL-I	29.73	26.56	33.71

Table 3: Performance of RoBERTa on different training set and testing set.

Pre-trained on → **Evaluated on ↓**	LogiQA-	ReClor-	RACE
NAIL	24.13	26.35	32.37
NAIL-E	20.20	28.41	33.60
NAIL-I	24.42	27.26	32.08

Table 4: Transfer learning results. Pre-trained on existing datasets and then fine-tuned on NAIL.

To investigate the effect of NAIL, NAIL-E and NAIL-I, we train a RoBERT $_{\text{LARGE}}$ model on NAIL, NAIL-E and NAIL-I separately and evaluate on the corresponding set of all tests (Table 3). When trained on NAIL-I and evaluated on NAIL-E, the model achieves the best accuracy 38.5%, showing that human imitations can significantly help models learn logic from natural text. When trained on NAIL-E and evaluated on NAIL-E, the model achieves the worst accuracy 22.9%. The possible reason is that the amount of data of NAIL-E is too small to train a deep model.

What are the results if the model is first trained on the existing reading comprehension dataset (LogiQA, ReClor and RACE are adopted), and then fine-tuned on the NAIL? To answer this question, we conduct a set of transfer learning experiments[4](Table 4). Overall, we observe that if models are first trained on RACE, they could obtain a significant improvement, since RACE is a large-scale dataset where models can learn some general reasoning ability. While interestingly, if models are first trained on ReClor and LogiQA, it is even close to random guess. We attribute this to the fact that, naïve logical reasoning is a typical category of reasoning which can be significantly distinguished from others. We believe that completely different categories of logical reasoning deserve to be explored separately since they may have different forms or strategies of reasoning, especially when training data is insufficient.

We further analyze the model performance with respect to different types of naïve logical reasoning (Figure 4). We find that language models perform well on set operation problems, while struggle on matching and ordering. We think that language models can provide good representation of set object, even if models do not really reason derived from the context.

5 DISCUSSION

We have some ideas for future directions of model design to solve NAIL. The first is to find the breakthrough of the question, i.e., the more certain a condition is to start from, the less possibilities it yields. For real/fake mixture, we also need to promptly identify two atomic statements that produce contradictory, that is, the two statements can not be both true(there must be one false) or both false(there must be one true), which is often the breakthrough in solving the problem.

The second is, tables and bipartite graphs can help to clarify the correspondence of given elements and conditions.

Another direction is allowing models to better utilize elimination. Models can perform forward elimination, i.e., according to context, every time we draw a definite conclusion, we can retain options with logical consistency and exclude those options that do not meet the conclusion.

6 RELATED WORK

There is abundant datasets in reading comprehension, which could facilitate the development of the field. MCTest (Richardson et al., 2013) is a multiple-choice reading comprehension dataset that contains 500 fictional stories and 2k questions. Rajpurkar et al. (2016) proposes the first large-scale reading comprehension dataset SQuAD (100k+ questions), where the answer to each question is a span of text from the passage. To push the development of reading comprehension, more datasets requiring more complicated reasoning types are introduced, such as multi-document datasets (Joshi et al., 2017; Dunn et al., 2017), multi-hop reasoning datasets (Yang et al., 2018; Welbl et al., 2018;

[4]ReClor- and LogiQA- denotes that, when training on LogiQA and ReClor, we removed all samples categorized to naïve logical reasoning, because they are included in NAIL.

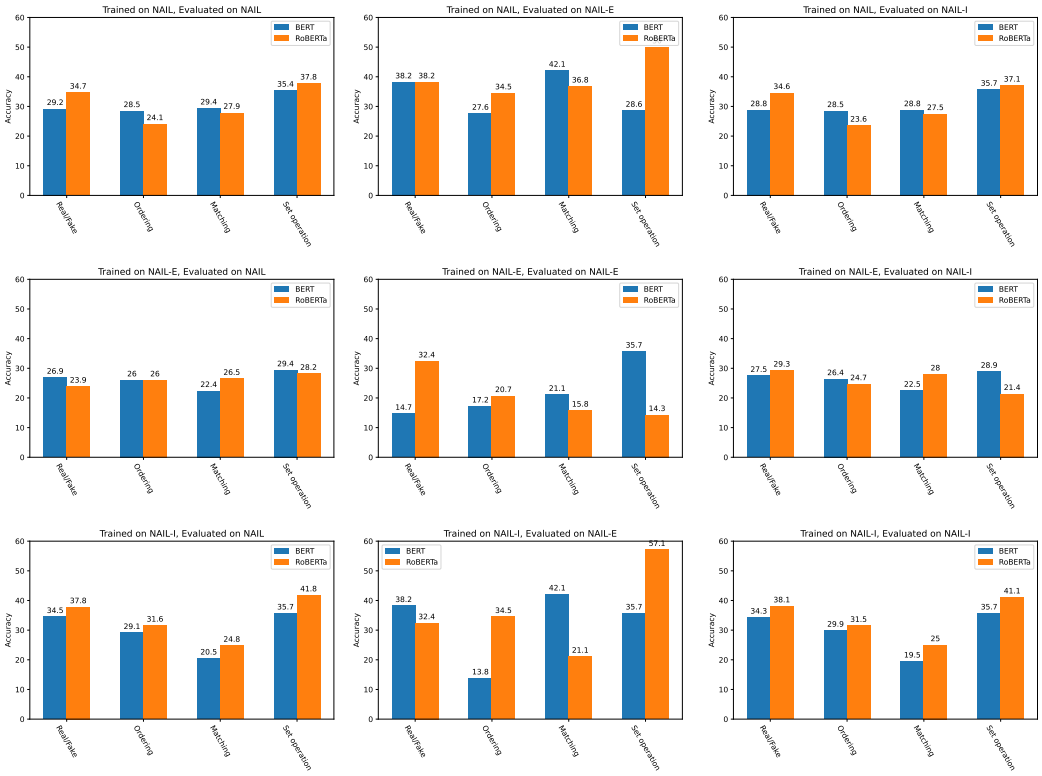

Figure 4: Accuracy against the reasoning type based on different training set and testing set.

Talmor & Berant, 2018), numerical discrete reasoning dataset (Dua et al., 2019) and commonsense reasoning datasets (Mihaylov et al., 2018; Zhang et al., 2018; Huang et al., 2019a). However, these datasets can not test the logical reasoning ability of the models. To fill this gap, LogiQA (Liu et al., 2020), ReClor (Yu et al., 2020) and LR-LSAT (Wang et al., 2021) was proposed.

LogiQA is collected from National Civil Servants Examination consisting of 8,678 questions and ReClor is extracted from Law School Admission Test consisting of 6,138 questions and LR-LSAT is sourced from Law School Admission Test, consisting of 4545 samples. These datasets require logical reasoning to answer. However, from human experience, there are different forms of reasoning strategy in answering different logical questions. We believe that completely different categories of logical reasoning deserve to be explored separately (Rudinger et al., 2020). Different from these these datasets, we inductively define a typical class of logical reasoning, named naïve logical reasoning, and then create a new benchmark targeting the task, named NAIL. Compared with LogiQA, Reclor and LR-LSAT, NAIL focus on the more fine-grained logical reasoning type (naïve logical reasoning).

There have been many datasets extracted from human examinations, such as LogiQA and ReClor mentioned above. Besides, RACE dataset (Lai et al., 2017) is collected from English exams for middle and high school Chinese students. ARC dataset (Clark et al., 2018) consists of 7,787 science exam questions drawn from a variety of sources.DREAM (Sun et al., 2019) is dialogue-based multiple-choice reading comprehension dataset collected from English as a Foreign Language examinations which contains 10,197 questions for 6,444 dialogues.

7 CONCLUSION

We introduce a more fine-grained logical reasoning, naïve logical reasoning, and propose a new large-scale benchmark NAIL targeted for learning and evaluating models' capabilities towards naïve logical reasoning. NAIL is source from standardized exams and human imitation. Preliminary results show that there is still a long way to go to equip deep models with true logical reasoning capability.

REFERENCES

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*, 2019.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.

Paul R Halmos. Naive set theory. 2017.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2391–2401, Hong Kong, China, November 2019a. Association for Computational Linguistics. doi: 10.18653/v1/D19-1243. URL https://www.aclweb.org/anthology/D19-1243.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2391–2401, 2019b.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 785–794, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1082. URL https://www.aclweb.org/anthology/D17-1082.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning, 2020.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

John McCarthy et al. *Programs with common sense*. RLE and MIT computation center, 1960.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018.

Allen Newell and Herbert Simon. The logic theory machine–a complex information processing system. *IRE Transactions on information theory*, 2(3):61–79, 1956.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://www.aclweb.org/anthology/D16-1264.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 193–203, 2013.

Rachel Rudinger, Vered Shwartz, Jena D Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A Smith, and Yejin Choi. Thinking like a skeptic: Defeasible inference in natural language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pp. 4661–4675, 2020.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231, 2019.

Alon Talmor and Jonathan Berant. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 641–651, 2018.

Siyuan Wang, Zhongkun Liu, Wanjun Zhong, Ming Zhou, Zhongyu Wei, Zhumin Chen, and Nan Duan. From lsat: The progress and challenges of complex reasoning. *arXiv preprint arXiv:2108.00648*, 2021.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302, 2018.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations (ICLR)*, April 2020.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. Record: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv:1810.12885*, 2018.

A IMPLEMENTATION DETAILS

We re-implement the rule-based methods following LogiQA (Liu et al., 2020). For pre-trained models, we modify the code of Transformers of HuggingFace (Wolf et al., 2020) to implement them on NAIL. We take the off-the-shelf model BERT and RoBERTa for NAIL, and Chinese BERT and Chinese RoBERTa (Cui et al., 2019) for Chinese NAIL. All hyper-parameters are selected by the model performance on the development sets.

B RESULTS ON VALIDATION SET

Trained on → **Evaluated on ↓**	NAIL	NAIL-E	NAIL-I
NAIL	34.63	24.91	31.98
NAIL-E	43.30	37.36	32.97
NAIL-I	34.19	24.28	31.93

Table 5: Performance of RoBERTa on different training set and validation set.

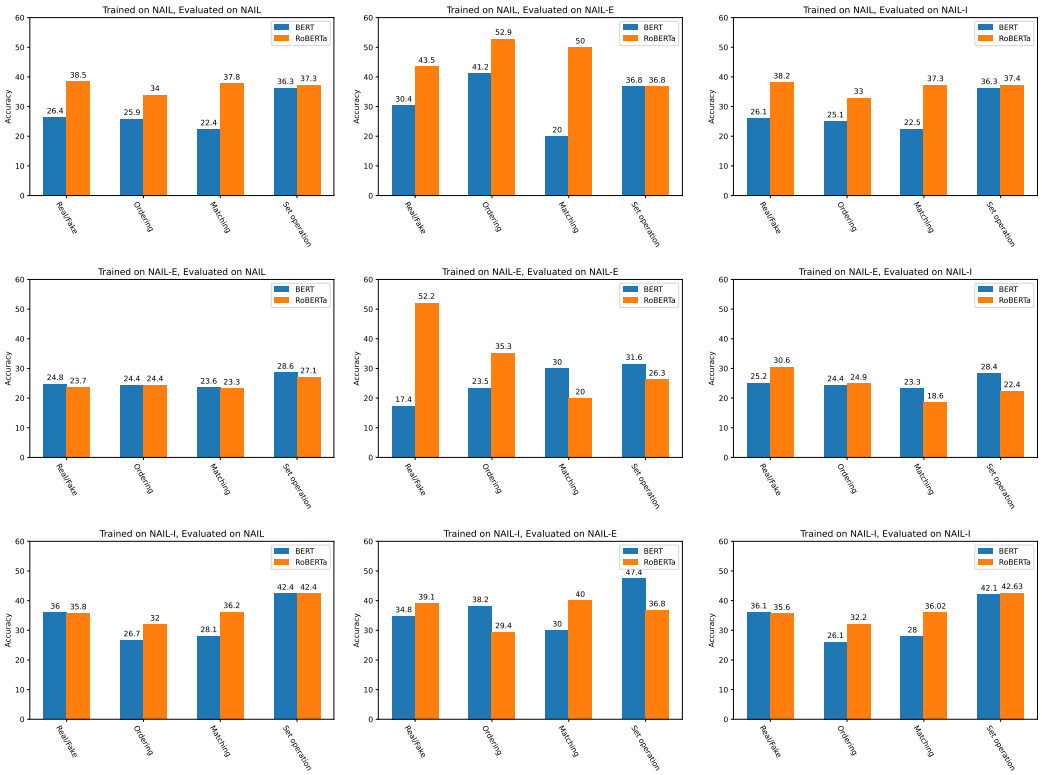

Figure 5: Accuracy against the reasoning type based on different training set and development set.