Abstract

Within psychology, philosophy, and cognitive science, *theory of mind* refers to the cognitive ability to reason about the mental states of other people, thus recognizing them as having beliefs, knowledge, intentions and emotions of their own. In this project, we construct a natural language inference (NLI) dataset that tests the ability of a state of the art language model, RoBERTa-large finetuned on the MNLI dataset, to make theory of mind inferences related to knowledge and belief. Experimental results suggest that the model struggles with such inferences, including after attempts for further finetuning.

1 Introduction

In this project, we examine to what extent the language model RoBERTa-large [1] finetuned on the MNLI dataset [2] can recognize inferences that involve *theory of mind*. The NLP task of natural language inference (NLI) is the task of predicting, given two sentences, a premise $X$ and a hypothesis $Y$, whether $X$ implies $Y$, contradicts $Y$, or is neutral with respect to $Y$ [3]. Theory of mind, also known as *folk psychology* within philosophy [4] or *intuitive psychology* in artificial intelligence [3], is a term that describes the human ability to reason about other humans’ mental states, thus recognizing them as having beliefs, knowledge, intentions and emotions of their own. Our theory of mind allows us to intuitively recognize entailments related to human mental states. Within theory of mind, our focus here is on verbs that describe epistemic mental states, like to know, to think, and to see. For an example of a theory of mind inference, note that the sentence *John knows that Ann thinks that there is milk in the fridge* entails that *Ann thinks that there is milk in the fridge* but not that *John thinks that there is milk in the fridge*. However, the sentence *John thinks that Ann knows that there is milk in the fridge* does imply *John thinks that there is milk in the fridge*. See Figure 1 for a graphical illustration.

Recent large transformer based language models have achieved impressive results in benchmarks like GLUE [5] and SuperGLUE [6], which aim to test natural language understanding, suggesting that larger models have an improved language understanding capacity. At the same time, recent work has exposed the limitations of such models on specially constructed NLI datasets [7, 8, 9, 10]. This project continues the latter line of work by introducing a new, automatically generated, dataset that tests language models understanding in theory of mind reasoning. Beyond offering a novel type of NLI test, we believe that the dataset offers an interesting challenge of linguistic understanding and reasoning. Theory of mind reasoning is considered an important developmental test within cognitive science and philosophy of mind, used for evaluating the cognitive capacities of young children and non-human animals [4]. According to this line of research, the ability of a thinking system to accurately represent the thoughts of a different system (which might deviate from the first system’s ‘world knowledge’) is an important mark of intelligence.

Our dataset contains three types of tests: the first type, intra-personal tests, involves reasoning about the mental states of a single agent. The second type, inter-personal tests, involves the mental states of multiple agents. Examples of such tests are given in Figure 1. The third type, inference reasoning,
John knows that Ann thinks that there is milk in the kitchen.

John thinks that there is milk in the kitchen.

Ann thinks that there is milk in the kitchen.

John thinks that Ann knows that there is milk in the kitchen.

Figure 1: Theory of mind entailments and non-entailments involving two agents and the epistemic states know and think.

Involves the ability to recognize other agents as making inferences. If $X$ entails $Y$, and you know that Bob believes that $X$, then (all things held equal), it is reasonable to conclude that Bob believes $Y$.

An important semantic distinction that any theory of mind reasoner (whether human or not) has to internalize is the distinction between factive mental state verbs and non-factive ones [11]. Factive verbs, like to know and to see, are assumed to accurately represent the facts in common use, or, to use linguistic terms, are assumed to presuppose their complements: if you know $X$, then $X$ is true. Non-factive verbs, like to think and to believe, do not come with such a presupposition: if you believe $X$, it does not imply that $X$ is true. Factivity plays an important role in the first two types of theory of mind reasoning tasks. The entailments and non-entailments in Figure 1 are largely explained by the factivity of the verb to know and the non-factivity of to think, respectively.

2 Related Work

Work on theory of mind inferences in language models builds on a diverse family of existing research fields:

Cognitive Science. Within cognitive science and philosophy of mind, the ability to ascribe beliefs to others, including false beliefs (requiring the awareness that agents can misrepresent reality) is seen as an important mark for developmental mental capacities [4]. Starting with [12], a large literature examines the ability of young children (years 3-5) in tasks that involve ascribing beliefs to others (the false belief task), and even ascribing beliefs among non-human animals, including primates [13] and corvids [14]. Within the intersection of cognitive science and NLP, [15] and [16] explore the ability of language models to do theory of mind reasoning through the task of question answering (QA). In particular, [16] argue that state of the art models for QA fail on carefully structured theory of mind tests, a result that is consistent with our results. In this project, we focus on the task of textual entailment, not QA.

Symbolic AI. Historically, work related to reasoning about the epistemic state of multiple agents has played an important role within symbolic artificial intelligence. In particular, epistemic logic is a formal system that logically allows to represent and reason about the knowledge states of agents [17, 18]. Epistemic logic contains the operator $K_a(p)$, representing the fact that agent $a$ knows the proposition $p$. Applications of epistemic logic include game theory [19], distributed computer systems [18] and AI planning [20]. Epistemic logic will be used in this project to schematize theory of mind inferences.
Linguistics and NLP. Within semantics, pragmatics, and philosophy of language, the ability to track the mental states of the participants of a conversation is assumed as a key ingredient in meaningful linguistic communication [21, 22, 23]. Theory of mind reasoning is therefore presupposed in those fields.

More specifically, the question we focus on is related to the NLP task known as speaker commitment or event factuality. In that task, an NLP model has to predict to what extent a speaker is committed to the complement of a sentence [24, 25]. For example, the model has to predict that the speaker of the sentence Michael knows that there is milk in the fridge is committed to the claim that there is milk in the fridge. The task of speaker commitment builds on extensive work in linguistics and formal semantics studying the presupposition behaviour of certain predicates, including epistemic verbs like to know [26, 11, 27]. Speaker commitment is related both to the task of presupposition inference (see [7]) and to theory of mind inferences, since it includes mental state verbs (like to know and to believe.) However, existing datasets used for speaker commitment tend to focus on complicated linguistic structures (involving negations, conditionals, and modals) with an emphasis on presupposition behaviour (see [24, 25]). Our dataset, on the other hand, involves simple linguistic structures, with a focus on complicated multi-agent scenarios.

Stress testing NLI models. Finally, this project is related to a recently growing literature that stress tests state of the art natural language inference models, using out of domain data and small manipulation on existing datasets. Manipulations include testing NLI models only on hypotheses [9], making small lexical changes on a single word in the example [28], and inducing spelling errors [29]. NLI tasks on specific out-of-domain datasets include datasets for defeasible reasoning [30], sentences with multiple quantifiers [31], entailment with conjunctions [32], the transitivity of the entailment relation [33], entailments involving event veridicality [24], and inferences involving presuppositions and implicatures [7]. An interrelated important line of work involves constructing NLI datasets that challenge the proposed entailment heuristics that state of the art models employ [29, 8]. In particular, the word overlap or subsequence heuristic (predict entailment if the hypothesis is a subsequence of the premise), which [29, 8] explore, will be relevant to our tests as well. Our dataset is generated by a very simple syntactic manipulation that can be easily reproduced for any type of dataset with minimal effort.

3 Approach

To evaluate our theory of mind dataset, we use the model RoBERTa-large finetuned on the MNLI dataset, available via Huggingface. The MNLI dataset contains 433K crowd-sourced and labeled examples of premise hypothesis pairs, from multiple genres, including examples from written and spoken sources [2]. RoBERTa-large finetuned on the MNLI dataset achieves a score of 0.908 on the MNLI test set [1].

The epistemic verbs that appear in the dataset are the factive verbs {know, understand, recognize, see, remember, learn} and the non-factives {believe, think, suspect, assume}. Since we are going to evaluate the model on examples containing these verbs, it is worth checking how much exposure the model had to them. Figure 2 plots the occurrences of these verbs in the MNLI training set (393K pairs of premise-hypothesis). The distribution ordinarily matches the general frequency of these verbs in English, according to the Oxford English Corpus, with significant representation of both the factives know (≈ 50K examples) and see (≈ 30K), and the non-factive think (≈ 25K). We note, however, that the high occurrence of the verb know may arise from its use as a discourse marker in English, and not necessarily from a strict epistemic use (see [34] for discussion).

Another possible worry is that the MNLI training dataset is skewed towards a particular label given a particular type of verb. Table 1, which compares the labels of MNLI training examples with the occurrences of factive and non-factive verbs in those examples, suggests that the training set is balanced. Given the above information, we hypothesize that the MNLI training dataset contains enough data to learn the basic semantic function of the verbs know, see and think and that therefore it

---

1See https://huggingface.co/transformers/ for the package and https://huggingface.co/roberta-large-MNLI for the model.

is meaningful to ask whether the model learns the function of these verbs (and similar verbs) in more complicated, theory of mind, inferences.

The new dataset only contains the labels entailment (label 1) and non-entailment (label 0), since the distinction between neutral and contradiction can be sometimes more difficult to draw. In what follows, we combine the labels contradiction and neutral to the single label non-entailment.

### 3.1 Dataset

For the purpose of describing the dataset we created, we follow the notation of epistemic logic and use $K_a$ (a knows that...) for an arbitrary factive verb from the set \{knows, sees, learns, understands, recognizes, remembers\}, $B_a$ (a believes that...) for an arbitrary non-factive verb from the set \{believes, thinks, assumes, suspects\}, and $V_a$ (a [some epistemic verb] that...) for an arbitrary (factive or not) epistemic verb. The subscript $a$ stands for a name $a$ taken randomly from a set of most common male and female names in the USA.\(^3\) All epistemic verbs are in singular, present, third-person form, unless stated otherwise.

To automatically create examples, we follow the following procedure. We randomly pick a premise $X$ from the SNLI dataset [35]. We then generate a new sentence by appending a given premise $X$ to a construction like Michael sees that or Ann thinks that (lower-casing the appropriate letters in $X$). For example, given the SNLI premise:

A black race car starts up in front of a crowd of people.

We generate the sentence:

Michael knows that a black race car starts up in front of a crowd of people.

We generate the sentence:

Michael knows that a black race car starts up in front of a crowd of people.

The latter sentence is schematized as $K_m(X)$. In the same fashion, more complicated sentences can be constructed, such as $K_m B_m(X)$. Linguistically, a sentence of the form $K_m(X)$ is grammatically acceptable iff $X$ is a grammatically acceptable declarative sentence.

Each example in the dataset is a tuple of the form (premise, hypothesis, label), where either the premise or the hypothesis (or both) were manipulated according to the above mentioned procedure. To determine the label, we follow the following widely accepted principles in epistemology and formal semantics:

---

\(^3\)Source https://www.ssa.gov/OACT/babynames/decades/names1990s.html
F: $K_a(X) \rightarrow X$ \hspace{0.5cm} \text{Factive verbs imply their complements}

NF: $B_a(X) \nRightarrow X$ \hspace{0.5cm} \text{Non-Factive verbs do not imply their complement}

C: $((X \rightarrow Y) \text{ and } V_a(X)) \rightarrow V_a(Y)$ \hspace{0.5cm} \text{Closure under entailment}

The principle C encodes the assumption that epistemic verbs are \textit{closed under entailments}: agents are rational and make inference of their own [36]. All three principles are regularly assumed in the semantics of epistemic verbs (see e.g. [37]). These rules imply, for example, the label \textit{entailment} in the tuple $(B_a, K_b(X), B_a(X), \text{entailment})$. These principles are not meant to represent logical or scientific truths, rather intuitive, or folk, assumptions. After all, Bob might believe $X$ without realizing that $X$ implies $Y$, and John might have a false memory that $X$, even though \textit{remember} is a factive verb. Nevertheless, since the task of NLI aims to represent commonsense, and not just logical inference, we can expect a competent model to learn and follow these principles.

The dataset is divided into three theory of mind test categories: \textit{intra-personal} reasoning, \textit{inter-personal} reasoning, and \textit{inference reasoning}. Each category includes several templates, and each template includes 300 examples.

**Intra-personal** theory of mind reasoning involves the mental states of a single person. The premise \textit{Michael thinks he knows that there is milk in the fridge} does not entail the hypothesis \textit{Michael knows that there is milk in the fridge}, since, as part of our intuitive theory of mind reasoning, \textit{thinking that you know} does not mean that you actually know. To generate the special construction $B_m K_m(X)$, we randomly pick an element from the set \{\textit{believes he knows, thinks he knows, thinks he remembers, thinks he saw, believes he saw}\} (for male names). This category also includes the simple control tests tuples $(K_m X, X, \text{entailment})$ and $(B_m X, X, \text{non-entailment})$ to test the model’s base understanding of factivity. The templates are:

$\begin{align*}
(K_m X, X, \text{entailment}) & \quad (B_m X, X, \text{no-entailment}) \\
(B_m K_m X, K_m X, \text{no-entailment}) & \quad (K_m B_m X, X, \text{no-entailment}) \\
(K_m B_m X, B_m X, \text{entailment}) &
\end{align*}$

**Inter-personal** theory of mind reasoning involves the mental states of multiple persons. Examples are provided in Figure 1. The inter-personal templates include:

$\begin{align*}
(B_a B_b(X), B_a(X), \text{no-entailment}) & \quad (B_a B_b(X), B_b(X), \text{no-entailment}) \\
(B_a K_b(X), K_b(X), \text{no-entailment}) & \quad (K_a B_b(X), K_b(X), \text{entailment}) \\
(K_b K_a(X), K_a(X), \text{entailment}) &
\end{align*}$

By \textit{inference reasoning}, we mean the ability of recognizing others as able to make inferences. If we recognize that premise $X$ entails hypothesis $Y$, we should also recognize that \textit{Michael believes X} entails \textit{Michael believes Y}. This is the tuple $(B_m(X), B_m(Y), \text{entailment})$ (where $X$ entails $Y$). At the same time, the premise \textit{Michael believes X} does not entail that \textit{Ann believes Y}, even if $X$ entails $Y$. Unlike the former categories, which are related to the task of speaker commitment (see [24], [25]), this kind of test, as far as we know, is novel in the NLI literature. The inference reasoning templates are

$\begin{align*}
(B_a(X), B_a(Y), \text{entailment}) & \quad (K_a(X), K_a(Y), \text{entailment}) \\
(B_a(X), B_b(Y), \text{no-entailment}) & \quad (K_a(X), K_b(Y), \text{no-entailment}) \\
(Forget_a(X), Forget_a(Y), \text{no-entailment}) & \quad (See_a(X), Know_a(Y), \text{entailment})
\end{align*}$

The tuple $(Forget_a(X), Forget_a(Y), 0)$, which we call \textit{Forget non-closure}, tests the special non-closure of forgetting: forgetting $X$ does not imply forgetting $Y$, even if $X$ implies $Y$. The tuple $((See_a(X), Know_a(Y), 1))$ which we call \textit{Sensory closure}, tests the intuition that knowledge is the most general factive state: if you see/recognize/realize $X$ and $X$ implies $Y$, you (generally) know $Y$ (see [38]). In this category, we pulled $(X, Y, \text{entailment})$ examples from the SNLI dataset. This category therefore also includes the control test $(X, Y, \text{entailment})$ to make sure that the model reliably recognizes $X$ as entailing $Y$.

\footnote{Formal reason: by combining $F$ and $C$ we get $((K_b X \rightarrow X) \text{ and } B_a(K_b X)) \rightarrow B_a(X)$. Therefore, the premise $B_b K_b X$, implies $B_a X$.}
4 Experiments

We feed the examples into three models: the RoBERTa-large finetuned on the MNLI dataset, and two further finetuned models on custom datasets we create, using the Huggingface transformers package.5 We use accuracy as an evaluation method. All results are in tables 2-5.

4.1 RoBERTa-Large-MNLI without further finetuning

We start by testing the dataset on the RoBERTa-large-MNLI model, as is from Huggingface. We tested this particular model on two types of source sentences X: X taken from SNLI premises (called in the tables RoBERTa-large-MNLI SNLI premises) and X taken for a list of shorter sentences (called in the tables RoBERTa-large-MNLI short sentences).

Results. The first two control columns of table 2 suggest that the model treats all verbs as factive verbs. The control test accuracy for non-factive mental states is 0.006. A possible explanation is that the model uses a word overlap heuristic predicting entailment (recognizing the overlap between \(B_x X\) and X), while ignoring the preceding verb \(B_x\). All high accuracy results are compatible with this heuristic. To test the assumption that the model is sensitive to the amount of word overlap, we further tested the templates on 400 shorter sentences (average length of 6 words, compared to 13 of the SNLI premises), taken from an ESL resources website.° The model’s accuracy improves, but not significantly. Furthermore, we add another template, explicitly adding a defeater modifier to non-factive verbs (i.e. wrongly, falsely, incorrectly thinks that X), creating the Anti-factive test (in table 2). Surprisingly, even here, the model performed below chance for the SNLI premises. Since the accuracy on the non-factive control is so low, the more complicated tests in tables 2 and 3 are uninformative for this model.

For the inference reasoning task (table 4), as expected, the model is performing well on the control (acc. 0.923), but remaining results show that the model treats modified \((X, Y)\) pairs as entailments under all modifications, thus failing to recognize that the verbs refer to different agents. For instance, the template \((B_x X, B_y Y, 0)\) has 0.27 accuracy.

4.2 RoBERTa-Large-MNLI with finetuning training set 1

Since the available RoBERTa-large-MNLI model performs poorly on the control tests (the template \(B_x X, X, 0\) in table 2), we try two approaches for further finetuning the model. Instead of randomly dividing the dataset into training, evaluation and tests sets, we create new custom training and evaluation sets. The idea behind this approach is to train the model on simple examples, and to see whether it improves on more complicated, held-out, ones. The first training set (training set 1) contained 3000 examples, with 1000 further examples for validation. Training set 1 had the following distribution of templates:

- 50%: \((B_x X, X, 0)\)
- 27%: unchanged SNLI examples
- 23%: \((K X, X, 1)\)

We finetune the model using the Trainer method of the Huggingface transformers package.7 We hypothesize that by learning the simple distinction between factive and non-factive verbs the model will improve performance on more complicated cases.

Results. The results of finetuning with training set 1 appear in the table rows RoBERTa-large-MNLI finetune training set 1. The hypothesis was mistaken. The results show that the finetuned model internalized that non-factive verbs are non factive (with acc. 1.0 in the control tests of table 2). However, the model just seems to use the heuristic:

if the premise or the hypothesis contain a non-factive verb, predict non-entailment; if they contain a factive verb, predict entailment.

---

5See the supplementary code.

6Source: https://7esl.com/english-verbs/

7See https://huggingface.co/transformers/training.html#trainer for that method. We finetune for 5 epochs, evaluating each epoch on the accuracy of the evaluation set, and using the AdamW optimizer and its default hyper-parameters. The model reached above 0.94 accuracy on the evaluation set during training for both finetuning set 1 and set 2.
<table>
<thead>
<tr>
<th></th>
<th>$K_a X$</th>
<th>$B_a X$</th>
<th>Anti-Factive $X$</th>
<th>$B_a K_a X$</th>
<th>$K_a B_a X$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RoBERTa-large-MNLI</strong></td>
<td>1.0</td>
<td>0.006</td>
<td>0.46</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>SNLI premises</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RoBERTa-large-MNLI</strong></td>
<td>1.0</td>
<td>0.254</td>
<td>0.882</td>
<td>0.024</td>
<td>1.0</td>
</tr>
<tr>
<td>short sentences</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RoBERTa-large-MNLI</strong></td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.133</td>
<td>1.0</td>
</tr>
<tr>
<td>finetune training set 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RoBERTa-large-MNLI</strong></td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.006</td>
</tr>
<tr>
<td>finetune training set 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Intra-personal tests

<table>
<thead>
<tr>
<th></th>
<th>$B_a B_b X$</th>
<th>$B_a B_b X$</th>
<th>$K_a K_b X$</th>
<th>$K_a K_b X$</th>
<th>$B_a K_b X$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RoBERTa-large-MNLI</strong></td>
<td>0.143</td>
<td>0.0</td>
<td>0.873</td>
<td>0.993</td>
<td>0.023</td>
</tr>
<tr>
<td>SNLI premises</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RoBERTa-large-MNLI</strong></td>
<td>0.152</td>
<td>0.002</td>
<td>0.970</td>
<td>0.995</td>
<td>0.039</td>
</tr>
<tr>
<td>short sentences</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RoBERTa-large-MNLI</strong></td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.333</td>
</tr>
<tr>
<td>finetune training set 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RoBERTa-large-MNLI</strong></td>
<td>1.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>finetune training set 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Inter-personal tests

<table>
<thead>
<tr>
<th></th>
<th>$B_a X$</th>
<th>$B_a Y$</th>
<th>$K_a X$</th>
<th>$K_a Y$</th>
<th>$B_a X$</th>
<th>$B_a Y$</th>
<th>$K_a X$</th>
<th>$K_a Y$</th>
<th>Forget non-closure</th>
<th>Sensory closure</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RoBERTa-large-MNLI</strong></td>
<td>0.923</td>
<td>0.936</td>
<td>0.926</td>
<td>0.27</td>
<td>0.283</td>
<td>0.066</td>
<td></td>
<td></td>
<td>0.943</td>
<td></td>
</tr>
<tr>
<td>SNLI Control (X, Y, 1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RoBERTa-large-MNLI</strong></td>
<td>0.933</td>
<td>0.003</td>
<td>0.933</td>
<td>0.996</td>
<td>0.013</td>
<td>0.006</td>
<td></td>
<td></td>
<td>0.993</td>
<td></td>
</tr>
<tr>
<td>finetune training set 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RoBERTa-large-MNLI</strong></td>
<td>0.89</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>finetune training set 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Inference reasoning tests

<table>
<thead>
<tr>
<th></th>
<th>$X$</th>
<th>$B_a K_b X$</th>
<th>$K_a K_b X$</th>
<th>$B_a B_b X$</th>
<th>$V_a V_b X$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RoBERTa-Large-MNLI</strong></td>
<td>0.53</td>
<td>1.0</td>
<td>0.463</td>
<td>0.826</td>
<td>0.4333</td>
</tr>
<tr>
<td>SNLI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RoBERTa-large-MNLI</strong></td>
<td>0.0</td>
<td>0.91</td>
<td>0.0</td>
<td>1.0</td>
<td>0.3333</td>
</tr>
<tr>
<td>training set 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RoBERTa-large-MNLI</strong></td>
<td>1.0</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>training set 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Additional tests, all models
This heuristic explains the high and low accuracies the model is getting. For instance, the template $(B_a B_b X, B_a X, 0)$ in table 3 gets full accuracy presumably just because it involves non-factive verbs, not because of the interpersonal structure. This explains the failure of the template $(B_a K_b X, K_b X, 0)$, with acc. of 0.333, since it involves both factive and non-factive verbs. Moreover, this finetuned model fails on the inference reasoning tasks that involve non-factive verbs (the template $(B_a X, B_a Y, 1)$ has 0.003 accuracy and $(K_a X, K_b Y, 0)$ has 0.013 accuracy). These low results are explained by the above heuristic. To further test this heuristic, we created additional templates (table 5). The finetuned model on training set 1 has 0 accuracy on the template $(X, K_a X, 0)$: the model seems to predict entailment just because there is a factive verb in the hypothesis. The further tests in table 5 show that the model cannot handle multi-agent examples. The template $(V_a V_b X, V_b V_a X, 0)$, in which the order of the verbs is swapped, receives low accuracy on the original model and the first finetuned model.

4.3 RoBERTa-Large-MNLI with finetuning training set 2

In the second fine tuning set (training set 2), we try to teach the model both to distinguish between factive and non-factive mental states and between the mental states of different agents. We hypothesize that this will result in abandoning the simple heuristic that emerged from training set 1. We finetune the original RoBERTa-large-MNLI on training set 2, with the same finetuning approach and setting as the first finetuning (see footnote 7), but with the following distribution of examples:

- 33.3%: $(B X, X, 0)$
- 33.3%: unchanged SNLI examples
- 12%: $(K X, X, 1)$
- 10%: $(X, V_a X, 0)$
- 5%: $(V_a V_b X, V_b V_a X, 0)$
- 5%: $(V_a V_b X, V_b V_a X, 0)$

The last two templates aim to teach the model the difference between the mental states of different agents.

Results. The results of the second finetuning appear in table rows RoBERTa-large-MNLI training set 2. Like in the first finetuning, the second finetuned model internalizes the simple difference between factive and non-factive verbs (see the first three columns of table 2). Further, as expected, the model receives a perfect score on the template $(V_a V_b X, V_b V_a X, 0)$ in table 5, which the model was trained on. However, the model still struggles with intra-personal combinations of factive and non-factive verbs (acc. of 0.006 on the template $(K_a B_a X, B_a X, 1)$). With respect to inter-personal and inference tests, the model now seems to follow the simple heuristic predict non-entailment if two agents are mentioned. This heuristic explains the perfect scores of this model in tables 3,4 and 5, as well as the perfect failure (acc. 0) in templates like $(B_a K_b X, B_a X, 1)$ (table 5) and $(K_a X, K_b Y, 1)$ (table 4) which require predicting entailments.

5 Analysis

All the tested models seem to resort to simple heuristics in order to make theory of mind inferences. This observation is consistent with recent work exposing the heuristics employed by NLI models [8]. The original RoBERTa-large-MNLI treats all mental states as factive, rendering it incapable to perform meaningful theory of mind inferences. Further finetuning efforts easily fix this bias. But instead of internalizing the F, NF and C principles that generate the bulk of our templates, the finetuned models employ simple heuristics that are consistent with the data in the finetuned set, but not with a general, recursive, understanding of these principles.

6 Conclusion

Theory of mind inferences constitute just a small fraction of all the inferences we expect an NLI model to perform, but—due to their role in cognitive science, semantics, and pragmatics—they are arguably an important fraction. In this project, we have focused on the limitations of one particular model on just a handful of theory of mind inferences related to knowledge and belief. One might argue that by finetuning only on a small subset of carefully picked examples, it is not surprising that the finetuned models fail to learn all templates. We note, however, that no set of examples is going to cover all the templates that can be generated by the principles F, NF, and C. It would be therefore an interesting challenge to try and develop, with the right training data, a robust NLI model for general theory of mind reasoning.
References


