# AnaLog: Testing Analytical and Deductive Logic Learnability in Language Models

Ryb, S.; Giulianelli, M.; Sinclair, A.; Fernández, R.

**DOI**
10.18653/v1/2022.starsem-1.5

**Publication date**
2022

**Document Version**
Final published version

**Published in**
The 11th Joint Conference on Lexical and Computational Semantics

# AnaLog: Testing Analytical and Deductive Logic Learnability in Language Models

**Samuel Ryb**
Tufts University
samuel.ryb@tufts.edu

**Mario Giulianelli**
University of Amsterdam
m.giulianelli@uva.nl

**Arabella Sinclair**
University of Aberdeen
arabella.sinclair@abdn.ac.uk

**Raquel Fernández**
University of Amsterdam
raquel.fernandez@uva.nl

## Abstract

We investigate the extent to which pre-trained language models acquire analytical and deductive logical reasoning capabilities as a side effect of learning word prediction. We present AnaLog, a natural language inference task designed to probe models for these capabilities, controlling for different invalid heuristics the models may adopt instead of learning the desired generalisations. We test four language models on AnaLog, finding that they have all learned, to a different extent, to encode information that is predictive of entailment beyond shallow heuristics such as lexical overlap and grammaticality. We closely analyse the best performing language model and show that while it performs more consistently than other language models across logical connectives and reasoning domains, it still is sensitive to lexical and syntactic variations in the realisation of logical statements.

## 1 Introduction

Logical reasoning (Lakoff, 1970; MacCartney and Manning, 2007; Smith, 2020) is at the core of many downstream NLP tasks, such as dialogue and story generation (Fan et al., 2018; Welleck et al., 2019); narrative understanding and summarisation (Mostafazadeh et al., 2016; Vashishtha et al., 2020); question answering (Weber et al., 2019; Shi et al., 2021); relation extraction (Massey et al., 2015; Kassner et al., 2020; Yanaka et al., 2021); and visual comprehension (Suhr et al., 2017, 2019; Sethuraman et al., 2021). Because most of the current approaches to these tasks rely on pre-trained language models (LMs), it is essential to understand whether LMs can perform logical reasoning.

One way of verifying LMs' reasoning abilities is using a natural language inference (NLI) task (Dagan et al., 2005; Giampiccolo et al., 2007; Bowman et al., 2015; Bhagavatula et al., 2020; Rudinger et al., 2020). In NLI, an LM is given a premise and a hypothesis, and its task is to predict the logical relation between the two. Yet, LMs typically learn to solve NLI by using invalid heuristics, for example by extracting overlapping patterns between premises and hypotheses (McCoy et al., 2019), or by using specific lexical items and sentence grammaticality as simplistic predictors of entailment (Poliak et al., 2018).

In this paper, we examine whether pre-trained LMs rely solely on shallow heuristics, or whether they can use relevant reasoning abilities to make inferences. To do so, we develop a new NLI task, **AnaLog**,[1] that requires LMs to encode different logical reasoning patterns and we probe the behaviour of four masked and autoregressive LMs on this new dataset. Using interpretability measures, we find that, as a side effect of learning word prediction, all LMs under scrutiny have—to some extent—learned to encode information that is predictive of entailment relations.

We analyse the behaviour of the best performing model, BERT (Devlin et al., 2019), across the various inference categories present in AnaLog, finding that its reasoning abilities go beyond shallow heuristics and yield relatively consistent performance on deductive and analytical reasoning, as well as across reasoning domains (spatial and comparative) and logical connectives. Nevertheless, the model's behaviour within connectives varies, pointing out its sensitivity to lexical and syntactic variations in the realisation of logical statements.

## 2 Related Work

### 2.1 Learning Logic from Text

Recent work has explored which aspects of logical reasoning are statistically learnable from text. Examining how well LMs encode the semantics of

[1]The dataset is available at https://github.com/dmg-illc/analog



*Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 55 - 68
July 14-15, 2022 

logical connectives can give us insight into their reasoning capabilities, i.e., their ability to reach a conclusion from one or more statements.

Kim et al. (2019b) showed that BERT (Devlin et al., 2019) achieves 10% higher accuracy than humans on tasks that involve conjunctions. However, it has also been shown that LMs fail to encode the semantics of logical formulas (Traylor et al., 2021b) and struggle to differentiate between conjunction and disjunction (Traylor et al., 2021a), particularly in instances where the operands are noun phrases (Talmor et al., 2020), suggesting that the models find it difficult to understand the scope of the logical operator. It is also known that neural LMs have difficulty understanding argument order (Kassner et al., 2020), which is arguably a pre-requisite for any logical reasoning. Clark et al. (2020) and Tian et al. (2021) showed that RoBERTa (Liu et al., 2019), in contrast to BERT, performs well at encoding instructional texts that involve conditionals. Good performance on conditionals in LMs is surprising, since humans typically find reasoning about conditionals challenging due to the fact that it requires accommodating degrees of belief (Politzer, 2007). Finally, regarding universal quantification, which implicitly involves encoding a hidden conditional statement (e.g. $\forall x.P(x) \rightarrow Q(x)$), BERT's performance has been shown to vary substantially (Kim et al., 2019b; Tian et al., 2021).

Besides different logical connectives, some recent work has studied different types of reasoning domains. Kassner et al. (2020) showed that models such as BERT and RoBERTa struggle to encode the semantics of comparative reasoning phrases. Yet, Kim et al. (2019b) showed that BERT's performance is only 11% less than human performance on comparative reasoning tasks, and 10% less than human performance on spatial reasoning tasks.

Overall, there is a lot of variation in LMs' abilities to interpret different aspects of logical reasoning. We suspect that low performance stems from the fact that LMs are struggling to encode world knowledge, which is often required in NLI and logic datasets (Clark et al., 2007; Wang et al., 2018; Lauscher et al., 2020; Kassner et al., 2020; Ryb and Van Schijndel, 2021), while high performance may be due to extracting overlapping heuristics (Beall et al., 2019; McCoy et al., 2019), or to attending to shallow predictors such as the presence of specific words or sentence grammaticality (Poliak et al., 2018). We control for these factors in AnaLog.

### 2.2 Diagnostic Probing

A well established way of investigating what type of linguistic information is tracked by neural LMs is *diagnostic probing* (Ettinger et al., 2016; Adi et al., 2017; Belinkov et al., 2017; Conneau et al., 2018; Hupkes et al., 2018). Probing typically consists of extracting model representations, feeding them as input to a supervised classifier trained to predict a hypothesised linguistic property (e.g., the grammatical number agreement of the main verb of a sentence), and testing the probing classifier on a set of unseen representations. Good probing performance cannot directly be taken to indicate that the hypothesised linguistic property is tracked by the LM (Belinkov, 2021). It is thus common practice to compare the true probing performance of classifiers with performance on control representations (Zhang and Bowman, 2018; Tenney et al., 2018; Chrupała et al., 2020), tasks (Hewitt and Liang, 2019a), or datasets (Ravichander et al., 2021).

In this paper, we set up a careful evaluation procedure to interpret the performance of our probing classifier, by training it on increasingly small portions of training data, and comparing its performance in relation to two baselines.

## 3 Dataset Design and Construction

We extend the LAKNLI dataset (Ryb and Van Schijndel, 2021) and present AnaLog, an NLI dataset that explicitly targets different types of logical reasoning. The dataset contains a total of 24,000 items (see Table 2), where each item consists of a premise, a hypothesis, and their logical relation: *entailment* or *non-entailment*. Premises and hypotheses are generated from templates, using a restricted and carefully selected vocabulary. The templates and the vocabulary can be found in Appendices A.1 and A.2. The dataset is designed to contain a balanced distribution of logical connectives and reasoning categories. Examples are provided in Table 1.

### 3.1 Premises

Sentences in AnaLog are constructed from templates designed for specific logical connectives. For example:

(1) $N_1\ P_1\ N_2$ ***and*** $N_3$

A premise is constructed through filling a tem-

| | Premise | Hypothesis | |
|---|---|---|---|
| | | OVERLAP | NO-OVERLAP |
| AND | *Jennifer is in front of Elizabeth **and** Jennifer is to the north of Linda.* | → *Jennifer is in front of Elizabeth.*<br>↛ *Elizabeth is to the north of Linda.* | → *A person is behind some woman.*<br>↛ *A person is behind some man.* |
| OR | *Jennifer is to the north of Linda **or** is below Robert. Jennifer is not below Robert.* | → *Jennifer is to the north of Linda.*<br>↛ *Robert is below Jennifer.* | → *Some person is to the south of some woman.*<br>↛ *Some boy is to the east of a man.* |
| CON | ***If** Elizabeth is older than Jennifer **then** Linda is smaller than Jennifer. Elizabeth is older than Jennifer.* | → *Linda is smaller than Jennifer.*<br>↛ *Jennifer is smaller than Linda.* | → *A person is larger than some woman.*<br>↛ *A woman is arriving later than some boy.* |
| UNI | ***Every** director is to the west of Patricia. James is a director.* | → *James is to the west of Patricia.*<br>↛ *Patricia is to the west of James.* | → *Some woman is to the east of some man.*<br>↛ *Some woman is to the right of some man.* |

Table 1: Examples of premises and hypotheses for each of the logical connectives. Within the premises, connectives are **bolded** and spatial and comparative reasoning predicates are highlighted in blue and orange, respectively.

plate's slots with nouns and predicates. For instance, $N_1$ = *Patricia*, $N_2$ = *James*, $N_3$ = *Mary*, and $P_1$ = *is to the left of* would result in:

(2) *Patricia is to the left of James **and** Mary*

**Logical Connectives** AnaLog systematically distinguishes between the following four types of logical connectives in the premise:

- AND: conjunction (*and*)
- OR: disjunction (*or*)
- CON: conditionals (*unless*, *if*, *if then*, *only if*)
- UNI: universal quantification (*every*, *all*)

This is in contrast to both SuperGLUE (Wang et al., 2020) where the logical connectives vary between being positioned in the premise or hypothesis, and LogicNLI (Tian et al., 2021), where premises consist of multiple facts and rules and do not isolate logical connectives. LogicNLI premises may also feature negation, existential quantification, and equivalence. Since negation is often used as a heuristic to predict non-entailment in NLI tasks (McCoy and Linzen, 2019), we only include it within premises when absolutely necessary to asses LMs' understanding of a specific reasoning schema (such as disjunction and certain forms of conditionals). Existential quantification and equivalence are implicitly present in our hypotheses construction, as explained in Section 3.2.

**Nouns** The noun slots in our premise templates are filled with proper names, as this avoids possible confounding factors carried over by the semantics of common nouns. We choose the eight most frequent male and female first names according to the 1990 U.S. Census Bureau's Population Division. For the restrictor noun in universal quantification premises (e.g., *director* in the UNI premise in Table 1), we use the four most common nouns in COCA (Davies, 2010) which correspond to the category NOUN.PERSON in Wordnet (Fellbaum, 1998), do not begin with a vowel,[2] and are semantically compatible with our predicates. Selecting high frequency nouns ensures that LMs are not thrown off by infrequent occurrences, nor heavily influenced by specific lexical material. This enables LMs to output representations that are as stable as possible.

**Predicates** The predicates in our templates are also instantiated with a restricted vocabulary that limits interference with additional sorts of knowledge. We focus on two reasoning domains: *spatial* (3) and *comparative* (4) reasoning. We select pairs of spatial reasoning predicates from Kim et al. (2019a), such as *left-right* and *above-below*. To collect comparative reasoning predicates, we select pairs from the FraCaS project (Cooper et al., 1996), such as *smaller-larger* and *weaker-stronger*. Reasoning about these two types of predicates requires models to encode truth equivalent relationships, such as:

(3) $N_1$ **is above** $N_2 \iff N_2$ **is below** $N_1$

(4) $N_1$ **is stronger than** $N_2 \iff N_2$ **is weaker than** $N_1$

### 3.2 Hypotheses

Assessing whether a given hypothesis is entailed by a premise may require different kinds of reasoning. For example, some hypotheses follow purely on the basis of structural aspects, i.e., they can be derived by direct deduction on surface form: e.g., '*A **and***

[2] So that they are all compatible with the article *a*.

*B*' logically entails '*A*' as well as '*B*', as in (5-a).[3] Such hypotheses require *deductive reasoning*. In contrast, other cases of entailment go beyond manipulations at the level of surface form and instead rely on additional semantic knowledge, as in (5-c). Such hypotheses require *analytical reasoning*.

To test both types of reasoning, we generate entailment and non-entailment hypotheses for each type. For the example premise in (5), this results in the following four hypotheses, where $\rightarrow$ denotes an entailment, and $\nrightarrow$ a non-entailment relation:

(5) *Patricia is to the left of James* ***and*** *Mary*
   a. $\rightarrow$ *Patricia is to the left of James*
   b. $\nrightarrow$ *Mary is to the left of James*
   c. $\rightarrow$ *Some man is to the right of some other person*
   d. $\nrightarrow$ *Some man is older than some woman*

For AND, we randomly select one of the conjuncts to construct the entailed direct logical deduction hypotheses. That is, (5-a) could have also been *Patricia is to the left of Mary*. Details of the other connectives can be found in Appendix A.2 (Table 7).

AnaLog clearly distinguishes between deductive and analytical reasoning, which gives rise to a systematic distinction between hypotheses that exhibit lexical overlap and those that do not exhibit any overlap of content words (see examples in Table 1). Hence, in addition to isolating LMs' abilities to both deductively and analytically reason, this offers a way to control LMs' potential use of overlap-related heuristics, which have been shown to artificially inflate previous results on the NLI task (McCoy et al., 2019). We explain this distinction in more detail next.

**Overlapping Hypotheses** Overlapping hypotheses only consist of words reiterated from the premise. *Overlapping entailment* ($O^{\rightarrow}$) hypotheses are a direct logical deduction (5-a), which corresponds to the strictest case of premise overlap considered by McCoy et al. (2019). *Overlapping non-entailment* ($O^{\nrightarrow}$) hypotheses, in contrast, do not logically follow from the premise (5-b). We generate two types of $O^{\nrightarrow}$ hypotheses: grammatical instances $O^{\nrightarrow}_{G}$ such as (5-b) and ungrammatical instances $O^{\nrightarrow}_{UG}$, which correspond to an ungrammatical bag-of-words subset of the premise (e.g. '*and to left the of Patricia*').

While it may not be realistic to expect that LMs have had exposure to ungrammatical sentences during training—and hence that they will have learned to properly reason with them (i.e., to systematically classify them as non-entailment)—including ungrammatical instances allows us to test the strength of possible overlap-based heuristics: if LMs more frequently incorrectly assign the label *entailment* to ungrammatical cases that exhibit lexical overlap, then we can consider lexical overlap as a stronger heuristic than grammaticality.

**Non-Overlapping Hypotheses** Non-overlapping hypotheses are generated by replacing proper names with person-related hypernyms and replacing the predicate with its counterpart (e.g., *James* $\rightsquigarrow$ *some man, left* $\rightsquigarrow$ *right*).[4] We generate both *Non-Overlap entailment* ($NO^{\rightarrow}$) hypotheses (i.e., proper instances of analytical reasoning, such as (5-c)) and *Non-Overlap non-entailment* ($NO^{\nrightarrow}$) hypotheses, such as (5-d).

| | O | E | G | AND | OR | CON | UNI |
|---|---|---|---|---|---|---|---|
| $O^{\rightarrow}$ | ✓ | ✓ | ✓ | 1,500 | 1,500 | 1,500 | 1,500 |
| $O^{\nrightarrow}_{G}$ | ✓ | ✗ | ✓ | 750 | 750 | 750 | 750 |
| $O^{\nrightarrow}_{UG}$ | ✓ | ✗ | ✗ | 750 | 750 | 750 | 750 |
| $NO^{\rightarrow}$ | ✗ | ✓ | ✓ | 1,500 | 1,500 | 1,500 | 1,500 |
| $NO^{\nrightarrow}$ | ✗ | ✗ | ✓ | 1,500 | 1,500 | 1,500 | 1,500 |
| | | | | 6,000 | 6,000 | 6,000 | 6,000 |

Table 2: AnaLog dataset statistics. The dataset contains 24,000 items in total. Overlap (O), Entailment (E), and Grammaticality (G) are marked. For each category (numerical cell), half of the items are constructed with spatial, and half with comparative reasoning predicates.

## 4 Experimental Setup

### 4.1 Models

We probe four pre-trained Transformer (Vaswani et al., 2017) language models using AnaLog. To ensure a fair comparison, we use the `large` architecture size for all models, as available in the HuggingFace library (Wolf et al., 2020). We compare the following architectures:

**BERT** (Devlin et al., 2019) A Transformer-based LM pre-trained on masked language modeling and

[3] In this example, '*B*' is the implicit proposition *'Patricia is to the left of Mary'*.

[4] We minimize the risk of the probe memorizing facts in the dataset by choosing to not have 1-to-1 mappings of proper names to person-related hypernyms.

next sentence prediction, known for its high performance at sentence and token classification tasks, including NLI (Talman and Chatzikyriakidis, 2019).

**LUKE** (Yamada et al., 2020) A masked LM with an entity-aware self-attention mechanism, that builds upon the RoBERTa architecture (Liu et al., 2019). Using LUKE enables us to investigate the degree to which entity tracking can assist in solving logic-based NLI.

**StructBERT** (Wang et al., 2019) A masked LM based on BERT with additional word and sentence order training objectives. We expect StructBERT to provide insight on whether structural cues are useful in solving logic-based NLI.

**GPT-2** (Radford et al., 2019) An autoregressive Transformer-based LM which is known for its high performance across text-generation tasks, yet has not been frequently tested on NLI datasets. We are interested in how abstract representations built by an autoregressive LM compare to those built by masked LMs.

### 4.2 Probing Procedure

For each premise-hypothesis pair in AnaLog, we concatenate the text of the premise with that of the hypothesis and with the special sentence token from each LM's vocabulary.[5] We feed this text to the LM and extract the last layer's hidden activations corresponding to the special token; we take the activations to be the abstract representation of a premise-hypothesis pair. Repeating this procedure for all the items in AnaLog, we collect a dataset of representations, which we split into a training and a test set (see Section 4.3). We fit a binary logistic regression classifier[6]—as more powerful classifiers have been shown to produce unreliable results (Hewitt and Liang, 2019a)—to the training set, obtain predictions for the test set, and compute accuracy and baselined probing scores, as described in the next section.

### 4.3 Controlled Evaluation

Diagnostic probes are known for achieving high accuracy on linguistic tasks despite representations not necessarily encoding relevant linguistic information (Hewitt and Liang, 2019b; Belinkov, 2021). To address this issue, following the approach taken by Zhang and Bowman (2018), we measure probing performance as the difference between the classification accuracy of the probing classifier trained on the original dataset, and the accuracy of a baseline. We call this *baselined probing performance* (BPP), adopting the terminology proposed by Hewitt et al. (2021). To select the strictest baseline setup, we consider two aspects: 1) the amount of data, and 2) the type of data—i.e., controlled baseline representations obtained from the AnaLog dataset, on which the probe is trained.

**Partial Training Sets** We split AnaLog into a main training and testing set using an 80-20 split. To prevent overfitting of the probing classifier, we evaluate it by varying the quantity of data it is exposed to: we create partial training sets by sampling increasingly larger fractions of our main training set (1%, 2%, 4%, 6%, 8%, 10%, 12.5%, 25%, 50%, 100%), using an approach similar to that of Zhang and Bowman (2018). The testing set remains fixed, so that regardless of the split and baseline probe, we evaluate on a consistent set of sentences. All the resulting training sets and the testing set are balanced with respect to the two classification labels (entailment and non-entailment), logical connectives, reasoning predicates, and overlap vs. non-overlap.

**Baselines** We train the probing classifier on two baseline settings. For the *Scrambled* baseline, we scramble words in the premises and hypotheses separately, and train the probing classifier on their concatenation. Humans should achieve 50% accuracy on this version of the dataset because random word order impedes logical reasoning. For the *Random* baseline, we train the probing classifier on randomly initialised vector representations.

We consider these baselines as sufficient to ensure that entailment relations can only be predicted by using logical reasoning and not by exploiting dataset artifacts. For example, if the probes were solely learning associations between proper names and person-related hypernyms, the scrambled probe could suffice to achieve the same performance as the probe optimised on the original AnaLog testing set.

We train the probing classifier from scratch for each LM, training split, and baseline. As shown in

[5] For BERT and StructBERT, we prepend the `[CLS]` token; for GPT-2, we append the `<|endoftext|>` token; for LUKE, we append the `</s>` token.

[6] We use the scikit-learn implementation with default hyperparameters. We do not tune the hyperparameters to reduce the risk of overfitting to the collected representations, which would inflate the probing results. All logistic regression classifiers are trained until convergence.

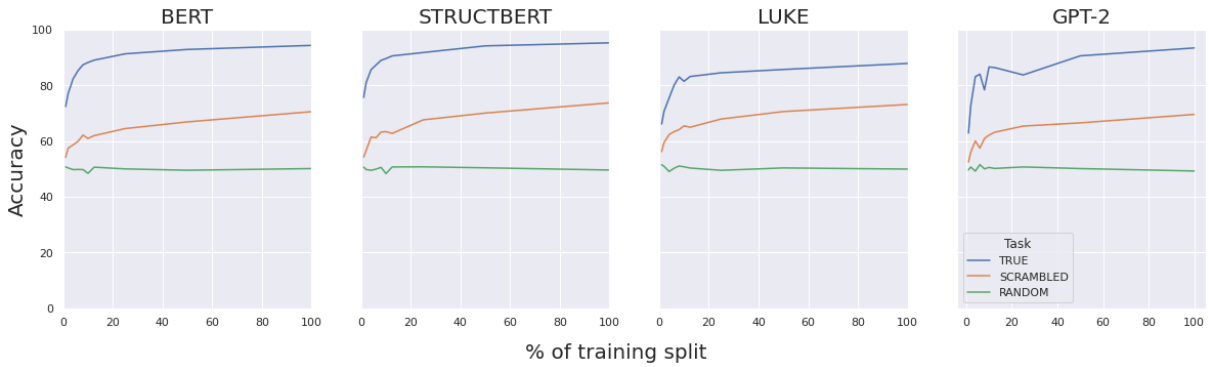



Figure 1: Accuracies of the original (*true*) vs. baseline (*scrambled, random*) probes for different training splits.

Figure 1, the *Scrambled* baseline achieves the highest accuracy (around 60%) across all LMs and training splits. The *Random* baseline achieves chance-level accuracy across LMs and training splits, confirming that the complexity of our probing classifier is appropriate for this task.[7] We therefore use *Scrambled* to compute BPP scores, as it yields the strictest (or most *selective*; Hewitt and Liang, 2019a) baseline setup.

## 5 Results across Models

All four LMs achieve positive average BPP scores: the average accuracy is above baseline by ca. 20 percentage points (see Figure 2). These overall results indicate that the LMs encode information that is predictive of entailment relations above and beyond simple heuristics which can be captured by a baseline. We also observe that the highest BPP scores are obtained at a relatively small training split size. This suggests training probes on more data can decrease their ability to extract the targeted linguistic features, and cause them to overfit on the dataset instead.

BERT and StructBERT are the best performing models with BPP scores ranging roughly between 15 and 40 (except for the smallest training split sizes). Their similar performance across all splits shows that StructBERT's explicit modelling of sentence and discourse structure does not produce more informative representations for our AnaLog task than BERT's simpler next word and next sentence prediction training objectives.

GPT-2's high standard deviation across splits (on average, 20.82) indicates a severe instability in its capacity to correctly encode logical reasoning cues. A closer look at GPT-2's performance shows that its representations are predictive of entailment relations when there is lexical overlap between premises and hypotheses, and of non-entailment relations when there is no lexical overlap. While GPT-2 is an autoregressive LM, as opposed to the other masked LMs, we are not certain that this factor is what causes this learning pattern. We leave exploring this further to future work.

Lastly, LUKE's performance, with an average score of 15.05, is significantly lower than that of the other three models ($t$-tests against BERT, StructBERT and GPT-2 yield $p$-values approaching zero), suggesting that its ability to track entities does not significantly help in solving logical deductions.

For the detailed results presented in the next sections, we focus on the model that achieves the highest BPP score with the lowest standard deviation. As can be seen in Figure 2, this model is BERT, probed with a classifier trained on 12.5% of the full training split.

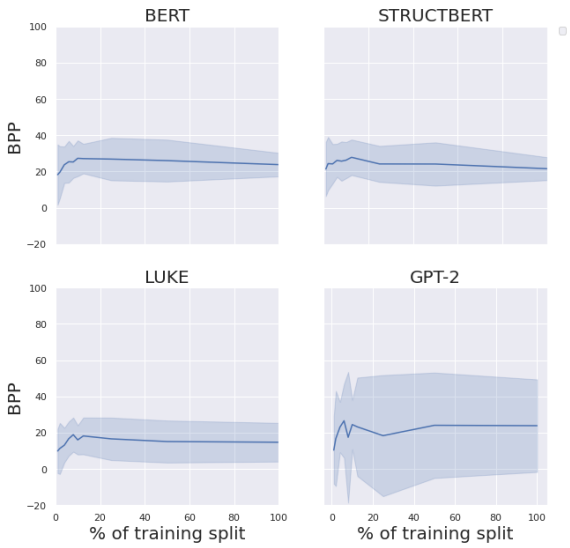
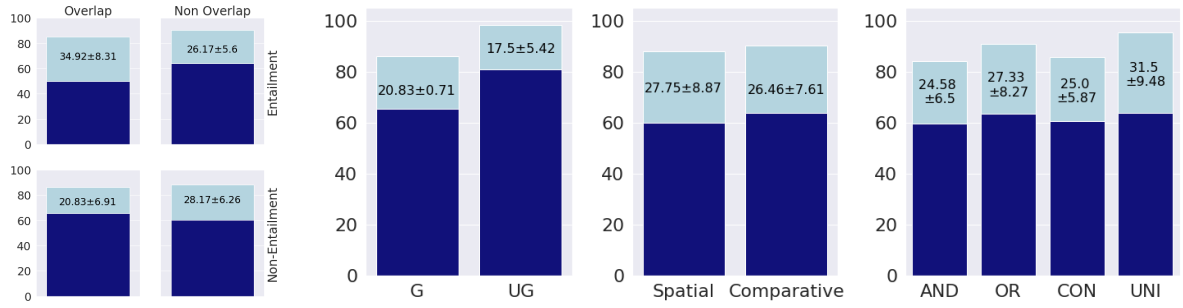



Figure 2: BPP scores for different training splits.

[7] We would have seen an accuracy greater than 50% for *Random* if the complexity of the classifier had been excessive.




(a) Overlap (all grammatical) (b) Grammaticality ($O^{\nrightarrow}$) (c) Reasoning domains (d) Logical connectives

Figure 3: BERT probing results across dataset categories. Overall bar height indicates accuracy, broken down by baseline accuracy (dark blue) and BPP score (light blue with superimposed average score and standard deviation).

## 6 Detailed Results with BERT

### 6.1 Solving Inference without Heuristics

We start by analysing the extent to which the performance of the best model, BERT, may be the result of exploiting heuristics unrelated to logical reasoning.

**Overlap** If lexical overlap were used as a heuristic to predict entailment, we would expect lower performance for *overlap-non-entailment* $O^{\nrightarrow}$ and *no-overlap-entailment* $NO^{\rightarrow}$ instances, where using the overlap heuristic yields incorrect predictions. This is not the pattern we observe. As shown in Figure 3a, accuracy is highest in these two cases. We see that $O^{\nrightarrow}$ items yield the lowest BPP scores and $NO^{\rightarrow}$ the highest (this difference is statistically significant and in principle compatible with the heuristics). However, there is no significant difference between no-overlap items with entailment vs. non-entailment labels. This indicates a lexical overlap heuristic is not prominently at play.

As pointed out in Section 3.2, the overlap vs. non-overlap distinction also corresponds to the contrast between direct deduction and analytical reasoning. We do not observe any significant differences in performance across these two reasoning types. More generally, the fact that BPP scores are positive across the board for overlapping and non-overlapping cases shows that the model is solving our logic-based NLI task by using information that goes beyond simple heuristic cues.

**Grammaticality** If a model were to judge entailment relations purely on the basis of grammaticality, we would expect it to wrongly predict *entailment* for $O^{\nrightarrow}_{G}$ (*overlap-non-entailment grammatical*) instances and correctly predict non-entailment for $O^{\nrightarrow}_{UG}$ (*overlap-non-entailment ungrammatical*). This is not what we observe: BPP scores are positive and not significantly different between $O^{\nrightarrow}_{G}$ and $O^{\nrightarrow}_{UG}$, which indicates grammaticality is not being used as a heuristic to predict entailment.

Finally, we find that performance on ungrammatical sentences is more unstable (standard deviation is almost 8 times higher than for $O^{\nrightarrow}_{G}$); this may be due to BERT producing noisier representations for out of distribution, partially ungrammatical, strings.

### 6.2 Consistency across Reasoning Domains

Having established that two plausible heuristics are not behind our probing results, we now turn to comparing reasoning domains. We have already seen that BERT's representations seem to be amenable to both deductive and analytical reasoning. We next hypothesize that if LMs can indeed reason logically, their performance should not be significantly affected by the specific choice of lexical items. We therefore compare the probes' performance on spatial vs. comparative reasoning predicates in AnaLog (see Figure 3c). We find no significant difference ($t = 0.442, p = 0.662$) in BPP scores across predicate types. This indicates that BERT's encoding of lexical semantic relations (in particular, antonymy) is stable across reasoning domains. This result is in line with the findings of Kim et al. (2019b), who show no substantial differences between spatial and comparative reasoning for BERT and humans.

### 6.3 Logical Connectives

Finally, we break down the results per logical connective. As can be seen in Figure 3d, BPP scores are positive and similar across operators, suggesting that BERT representations encode the semantics of logical connectives in a relatively stable way.

We observe the lowest BPP scores with conjunction and conditionals (in both cases significantly lower than UNI, $p < 0.05$). This is somewhat surprising, particularly for conjunction, given the previous results by Kim et al. (2019b) mentioned in Section 2.1. In the next section, we conduct two case studies to further examine whether there are specific linguistic phenomena linked to conjunction and conditionals that may be confusing BERT.

## 7 Analysis

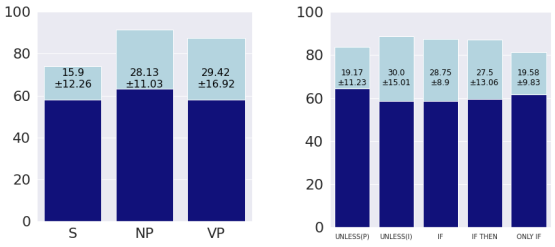
### 7.1 Case Study 1: Parsing Conjunction

In AnaLog, the arguments of a conjunction can be sentences (S), noun phrases (NP), or verb phrases (VP).[8] For example, the AND premise in Table 1 includes sentential conjuncts, while the one in example (5) features conjuncts that are NPs. We test two related hypotheses regarding aspects that may lead to lower performance in some of these conditions: $(i)$ We conjecture that, when the conjuncts are NPs or VPs, deducing information to the right of the conjunct may be more difficult because this involves parsing long-range dependencies. For example, in instances such as *David is to the left of John and Linda* → *Some girl is to the right of a boy*, predicting the entailment relation requires encoding syntactic and semantic information to both the left and right of the logical connective. $(ii)$ Consequently, we hypothesise that identifying the arguments of a conjunction may be easier for the model when these arguments are sentential rather than phrasal, since the former does not require parsing long-range dependencies; this would be compatible with the results by Talmor et al. (2020), who found that models struggle at making correct predictions when the conjunction is positioned between NPs.

Our two hypotheses, however, are not confirmed. On the one hand, we find no significant difference between left and right for any conjunct type (S, NP, and VP). This suggests that BERT's representations consistently encode information regardless of its position relative to the conjunction operator, which could be due to BERT's bidirectional training. On the other hand, as can be seen in Figure 4a, we observe that when the conjunction is positioned between sentences, the results are in fact significantly *worse* than when it is positioned between NPs or VPs.[9] Why this may be the case remains an open question that we leave for future work.

[8]These three types appear with equal frequency.

[9]All relevant $t$-tests yielded $p > 0.05$.




(a) Conjunction (b) Conditionals

Figure 4: BERT results within logical connectives.

### 7.2 Case Study 2: Types of Conditional

In this second case study, we investigate whether BERT's representations struggle to encode some types of conditionals more than others.[10] We expect to observe the highest performance for *if then* sentences, as BERT and RoBERTA reason well about modus-ponens (Clark et al., 2020). However, as shown in Figure 4b there is no significant differences between *if then*, *if*, and *unless(infix)*. The most challenging types are *only if* and *unless(prefix)*. We find that *unless(prefix)* is significantly outperformed by *unless(infix)*. This again shows that BERT is able to successfully encode relevant information to both the left and right of a connective.

## 8 Conclusions

We present a new NLI dataset, AnaLog, designed to test LMs' abilities to deductively and analytically reason. We choose diagnostic probing as an interpretability technique, and probe using AnaLog to inspect whether LMs acquire such logical reasoning abilities from text-based pre-training. We find that masked LMs, in particular BERT and StructBERT, can solve the inference task through encoding properties of both deductive and analytic logic, rather than solely relying on shallow heuristics such as lexical overlap and sentence grammaticality.

One main benefit of AnaLog is that it isolates different reasoning types, domains, and logical connectives, in order to gain a better understanding of which of these factors makes inference more challenging for an LM. We choose high frequency lexical items to ensure that the LMs' representations are as stable as possible, and not thrown off by surprising low frequency occurrences. We also use a fine-grained probing setup consisting of different

[10]The conditionals present in AnaLog are: *if*, *if then*, *only if*, *unless(prefix)*, *unless(infix)*; see Appendix A.2.

training splits and multiple baselines to ensure that probes are using relevant linguistic and logical information, rather than learning the dataset artifacts, to solve the task.

We perform an in-depth analysis of BERT's behaviour. Its overall stable performance is promising, though our case studies show some variance at the level of different natural language formulations of the same logical connective or their arguments as opposed to at higher reasoning levels. Overall, we think that BERT learns to encode approximations of the types of logical reasoning information necessary to solve AnaLog, although its sensitivity to surface forms can make these approximations inconsistent. While extending the AnaLog test set to also include lower frequency items may be helpful to ensure generalizability over noun and predicate relations (which we leave for future work), we hope that as it currently stands, AnaLog can be used as a benchmark to check whether LMs reason correctly by using elementary linguistic knowledge and logical semantics, as opposed to surface heuristics.

## Acknowledgements

We would like to thank the anonymous ARR and *SEM 2022 reviewers for their feedback and suggestions, as well as Ece Takmaz for her comments. Samuel Ryb and Arabella Sinclair worked on this project while affiliated with the University of Amsterdam. The project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819455).

## References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Jc Beall, Greg Restall, and Gil Sagi. 2019. Logical Consequence. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Spring 2019 edition. Metaphysics Research Lab, Stanford University.

Yonatan Belinkov. 2021. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, pages 1–13.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

Samuel R. Bowman, Christopher Potts, and Christopher D. Manning. 2015. Recursive neural networks can learn logical semantics. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pages 12–21, Beijing, China. Association for Computational Linguistics.

Grzegorz Chrupała, Bertrand Higy, and Afra Alishahi. 2020. Analyzing analytical methods: The case of phonology in neural models of spoken language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4146–4156.

Peter Clark, Phil Harrison, John Thompson, William Murray, Jerry Hobbs, and Christiane Fellbaum. 2007. On the role of lexical and world knowledge in RTE3. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 54–59, Prague. Association for Computational Linguistics.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3882–3890. International Joint Conferences on Artificial Intelligence Organization. Main track.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Robin Cooper, Richard Crouch, Jan van Eijck, Chris Fox, Josef Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, Steve Pulman, Ted Briscoe, Holger Maier, and Karsten Konrad. 1996. Using the Framework: The FraCaS Consortium. Technical report, FraCaS deliverable D-16.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Proceedings of the Machine Learning Challenges Workshop*, pages 177–190.

Mark Davies. 2010. The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, 25(4).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague. Association for Computational Linguistics.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher D. Manning. 2021. Conditional probing: measuring usable information beyond a baseline.

John Hewitt and Percy Liang. 2019a. Designing and interpreting probes with control tasks. 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)*, pages 2733–2743.

John Hewitt and Percy Liang. 2019b. Designing and interpreting probes with control tasks. 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)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Nora Kassner, Benno Krojer, and Hinrich Schütze. 2020. Are pretrained language models symbolic reasoners over knowledge? In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 552–564, Online. Association for Computational Linguistics.

Najoung Kim, Roma Patel, Adam Poliak, Alex Wang, Patrick Xia, R. Thomas McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019a. Probing what different NLP tasks teach machines about function word comprehension. In **SEMEVAL*.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019b. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

George Lakoff. 1970. Linguistics and natural logic. *Synthese*, 22(1/2):151–271.

Anne Lauscher, Olga Majewska, Leonardo F. R. Ribeiro, Iryna Gurevych, Nikolai Rozanov, and Goran Glavaš. 2020. Common sense or world knowledge? investigating adapter-based knowledge injection into pretrained transformers. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 43–49, Online. Association for Computational Linguistics.

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

Bill MacCartney and Christopher D. Manning. 2007. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 193–200, Prague. Association for Computational Linguistics.

Philip Massey, Patrick Xia, David Bamman, and Noah Smith. 2015. Annotating character relationships in literary texts. arXiv:1512.00728.

Richard T McCoy and Tal Linzen. 2019. Non-entailed subsequences as a challenge for natural language inference. *Proceedings of the Society for Computation in Linguistics*, 2(1):358–360.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448,

Florence, Italy. Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Guy Politzer. 2007. Reasoning with conditionals. *Topoi*, 26:79–95.

Alec Radford, Jeff Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI blog.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377.

Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675, Online. Association for Computational Linguistics.

Samuel Ryb and Marten Van Schijndel. 2021. Analytical, symbolic and first-order reasoning within neural architectures. In *Proceedings of the 2021 Workshop on Computing Semantics with Types, Frames and Related Structures*.

Muralikrishnna Sethuraman, Ali Payani, Faramarz Fekri, and James Kerce. 2021. Visual question answering based on formal logic. In *Proceedings of the 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 952–957.

Jihao Shi, Xiao Ding, Li Du, Ting Liu, and Bing Qin. 2021. Neural natural logic inference for interpretable question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3673–3684, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robin Smith. 2020. Aristotle's Logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Fall 2020 edition. Metaphysics Research Lab, Stanford University.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, Vancouver, Canada. Association for Computational Linguistics.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Aarne Talman and Stergios Chatzikyriakidis. 2019. Testing the generalization power of neural network models across NLI benchmarks. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 85–94, Florence, Italy. Association for Computational Linguistics.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2018. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. 2021. Diagnosing the first-order logical reasoning ability through LogicNLI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3738–3747, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aaron Traylor, Roman Feiman, and Ellie Pavlick. 2021a. AND does not mean OR: Using formal languages to study language models' representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 158–167, Online. Association for Computational Linguistics.

Aaron Traylor, Ellie Pavlick, and Roman Feiman. 2021b. Transferring representations of logical connectives. In *Proceedings of the 1st and 2nd Workshops on Natural Logic Meets Machine Learning*

*(NALOMA)*, pages 22–25, Groningen, the Netherlands (online). Association for Computational Linguistics.

Siddharth Vashishtha, Adam Poliak, Yash Kumar Lal, Benjamin Van Durme, and Aaron Steven White. 2020. Temporal reasoning in natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4070–4078, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. Superglue: A stickier benchmark for general-purpose language understanding systems.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, and Luo Si. 2019. Structbert: Incorporating language structures into pre-training for deep language understanding.

Leon Weber, Pasquale Minervini, Jannes Münchmeyer, Ulf Leser, and Tim Rocktäschel. 2019. NLProlog: Reasoning with weak unification for question answering in natural language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6151–6161, Florence, Italy. Association for Computational Linguistics.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, and Kentaro Inui. 2021. Exploring transitivity in neural NLI models through veridicality. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 920–934, Online. Association for Computational Linguistics.

Kelly Zhang and Samuel Bowman. 2018. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361.

# Appendix

## A Dataset Construction Details

### A.1 Lexical Items

Tables 3 and 6 respectively, show the noun, spatial and comparative analytic reasoning phrases used in AnaLog.

| Name | Gender | % Freq. | Count |
|---|---|---|---|
| James | M | 3.318 | 4,840,833 |
| John | M | 3.271 | 4,772,262 |
| Robert | M | 3.143 | 4,585,515 |
| Michael | M | 2.629 | 3,835,609 |
| William | M | 2.451 | 3,575,914 |
| David | M | 2.363 | 3,447,525 |
| Richard | M | 1.703 | 2,484,611 |
| Charles | M | 1.523 | 2,221,998 |
| Mary | F | 2.629 | 3,991,060 |
| Patricia | F | 1.073 | 1,628,911 |
| Linda | F | 1.035 | 1,571,224 |
| Barbara | F | 0.98 | 1,487,729 |
| Elizabeth | F | 0.937 | 1,422,451 |
| Jennifer | F | 0.932 | 1,414,861 |
| Maria | F | 0.828 | 1,256,979 |
| Susan | F | 0.794 | 1,205,364 |

Table 3: Noun phrases. Source: 1990 U.S. Census Bureau's Population Division.

As mentioned in Section 3.1, for the restrictors of the universal quantification premises (i.e., the $\text{UNI}_N$ slot in the Table 7 template), we used the four most common nouns in COCA (Davies, 2010) which do not begin with a vowel, and that correspond to the category NOUN.PERSON in Wordnet (Fellbaum, 1998), ensuring grammaticality when used within our templates (see Table 4).

| Restrictor Noun | POS | Frequency |
|---|---|---|
| model | n | 191,448 |
| director | n | 158,028 |
| participant | n | 81,371 |
| soldier | n | 78,276 |

Table 4: $\text{UNI}_N$ restrictor noun entries. Source: Corpus of Contemporary American English. POS stands for Part of Speech.

We replace the nouns from Table 3 with lexical entries from Table 5 within *non-overlapping entailment* ($\text{NO}^{\rightarrow}$) and *non-overlapping non-entailment* ($\text{NO}^{\nrightarrow}$) sentences, to ensure that models (and probes) are not using non-linguistic heuristics when solving the inference task.

| Gender | Hypernyms |
|---|---|
| Female | a girl, some girl, some other girl, a woman, some woman, some person, a person |
| Male | a boy, some boy, some other boy, a man, some man, some person, a person |

Table 5: Noun hypernyms used within AnaLog.

### A.2 Premise Constructions

Premises are constructed according to different templates (see Table 7). Let N be some noun (e.g. Patricia, David ...) and P be some spatial or comparative reasoning predicate (e.g. *is to the right of*, *is younger than ...* ). We use the $\neg$ symbol to denote negation. See Table 8 for information pertaining to the Specificity.

## B Computing Infrastructure and Budget

Our experiments were carried out using a single GPU on a computer cluster with Debian Linux OS. The GPU nodes on the cluster are GPU GeForce 1080Ti, 11GB GDDR5X, with NVIDIA driver version 418.56 and CUDA version 10.1. The total computational budget required to perform all our experiments amounts to 15 hours.

| Spatial Reasoning | Comparative Reasoning |
|---|---|
| $N_1$ **is to the left of** $N_2$ $\iff$ $N_2$ **is to the right of** $N_1$ | $N_1$ **is smaller than** $N_2$ $\iff$ $N_2$ **is larger than** $N_1$ |
| $N_1$ **is on top of** $N_2$ $\iff$ $N_2$ **is below** $N_1$ | $N_1$ **is faster than** $N_2$ $\iff$ $N_2$ **is slower than** $N_1$ |
| $N_1$ **is to the north of** $N_2$ $\iff$ $N_2$ **is to the south of** $N_1$ | $N_1$ **is arriving earlier than** $N_2$ $\iff$ $N_2$ **is arriving later than** $N_1$ |
| $N_1$ **is in front of** $N_2$ $\iff$ $N_2$ **is behind** $N_1$ | $N_1$ **is stronger than** $N_2$ $\iff$ $N_2$ **is weaker than** $N_1$ |
| $N_1$ **is to the east of** $N_2$ $\iff$ $N_2$ **is to the west of** $N_1$ | $N_1$ **is younger than** $N_2$ $\iff$ $N_2$ **is older than** $N_1$ |

Table 6: Predicates and their reasoning categories.

| LC | Specificity | Premise | Overlap Entailment |
|---|---|---|---|
| AND | S | $N_1$ $P_1$ $N_2$ **and** $N_3$ $P_2$ $N_4$. | Random[$N_1$ $P_1$ $N_2$, $N_3$ $P_2$ $N_4$]. |
| AND | NP | $N_1$ $P_1$ $N_2$ **and** $N_3$. | Random[$N_1$ $P_1$ $N_2$, $N_1$ $P_1$ $N_3$]. |
| AND | VP | $N_1$ $P_1$ $N_2$ **and** $P_2$ $N_3$. | Random[$N_1$ $P_1$ $N_2$, $N_1$ $P_2$ $N_3$]. |
| OR | S | $N_1$ $P_1$ $N_2$ **or** $N_3$ $P_2$ $N_4$. Random[$N_1$ $\neg$ $P_1$ $N_2$, $N_3$ $\neg$ $P_2$ $N_4$]. | The non-negated non-selected random sentence. |
| OR | NP | P: $N_1$ $P_1$ $N_2$ **or** $N_3$. Random[$N_1$ $\neg$ $P_1$ $N_2$, $N_1$ $\neg$ $P_1$ $N_3$]. | The non-negated non-selected random sentence. |
| OR | VP | $N_1$ $P_1$ $N_2$ **or** $P_2$ $N_3$. Random[$N_1$ $\neg$ $P_1$ $N_2$, $N_1$ $\neg$ $P_2$ $N_3$]. | The non-negated non-selected random sentence. |
| CON | UNLESS Prefix | **Unless** $N_1$ $P_1$ $N_2$, $N_3$ $P_2$ $N_4$. $N_1$ $\neg$ $P_1$ $N_2$. | $N_3$ $P_2$ $N_4$. |
| CON | UNLESS Infix | $N_1$ $P_1$ $N_2$ **unless** $N_3$ $P_2$ $N_4$. $N_3$ $\neg$ $P_2$ $N_4$. | $N_1$ $P_1$ $N_2$. |
| CON | IF | $N_1$ $P_1$ $N_2$ Random[**if, when, even though**] $N_3$ $P_2$ $N_4$. $N_3$ $P_2$ $N_4$. | $N_1$ $P_1$ $N_2$. |
| CON | IF THEN | **If** $N_1$ $P_1$ $N_2$ **then** $N_3$ $P_2$ $N_4$. $N_1$ $P_1$ $N_2$. | $N_3$ $P_2$ $N_4$. |
| CON | ONLY IF | $N_1$ $P_1$ $N_2$ **only if** $N_3$ $P_2$ $N_4$. $N_1$ $P_1$ $N_2$. | $N_3$ $P_2$ $N_4$. |
| UNI | Each | **Each** $UNI_N$ $P_1$ $N_1$. $N_2$ is a $UNI_N$. | $N_2$ $P_1$ $N_1$. |
| UNI | Every | **Every** $UNI_N$ $P_1$ $N_1$. $N_2$ is a $UNI_N$. | $N_2$ $P_1$ $N_1$. |

Table 7: Syntactic templates for premises and their corresponding overlapping entailment hypotheses. The logical connectives (LC) are **bolded** within each premise. Specificity indicates the lexical representation and/or the position in which the LCs are used within premises.

| Specificity | Definition |
|---|---|
| S | Conjunction/disjunction is positioned between sentences. |
| NP | Conjunction/disjunction is positioned between between noun phrases. |
| VP | Conjunction/disjunction is positioned between verb phrases. |
| UNLESS Prefix | The logical conditional connective is denoted by the word *unless* prefixed to the premise. |
| UNLESS Infix | The logical conditional connective is denoted by the word *unless* within the premise. |
| IF | The logical conditional connective is denoted by the word *if*. |
| IF THEN | The logical conditional connective is denoted by the phrase *if ... then ...*. |
| ONLY IF | The logical conditional connective is denoted by the phrase *only if*. |
| Each | The universal quantifier is denoted by the word *each*. |
| Every | The universal quantifier is denoted by the word *every*. |

Table 8: Specificity definitions.