ImplicaTR: A Granular Dataset for Natural Language Inference and Pragmatic Reasoning in Turkish

Mustafa Kürşat Halat
Boğaziçi University, İstanbul
kursathalat@gmail.com

Ümit Atlamaz
Boğaziçi University, İstanbul
umit.atlamaz@bogazici.edu.tr

Abstract

We introduce ImplicaTR, a linguistically informed diagnostic dataset designed to evaluate semantic and pragmatic reasoning capabilities of Natural Language Inference (NLI) models in Turkish. Existing Turkish NLI datasets treat NLI as determining whether a sentence pair represents *entailment*, *contradiction*, or a *neutral* relation. Such datasets do not distinguish between *semantic entailment* and *pragmatic implicature*, which linguists have long recognized as separate inferences types. ImplicaTR addresses this by testing NLI models' ability to differentiate between *entailment* and *implicature*, thus assessing their pragmatic reasoning skills. The dataset consists of 19,350 semi-automatically generated sentence pairs covering *implicature, entailment, contradiction,* and *neutral* relations. We evaluated various models (BERT, Gemma, Llama-2, and Mistral) on ImplicaTR and found out that these models can reach up to 98% accuracy on semantic and pragmatic reasoning. We also fine tuned various models on subsets of ImplicaTR to test the abilities of NLI models to generalize across unseen implicature contexts. Our results indicate that model performance is highly dependent on the diversity of linguistic expressions within each subset, highlighting a weakness in the abstract generalization capabilities of large language models regarding pragmatic reasoning. We share all the code, models, and the dataset.[1]

1 Introduction

Natural Language Inference (NLI) tasks are generally designed as three-way classification problems between sentence pairs (Gubelmann et al., 2023). Given a sentence pair consisting of a premise (P) and a hypothesis (H), the task is to classify the relation between P and H as one of *entailment, contradiction,* or *neutral*. Some of the most commonly used NLI datasets such as SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) contain three way annotations of sentence pairs and recently Budur et al. (2020) translated both datasets into Turkish to create the combined NLI-TR dataset. Although these NLI datasets have been useful in testing the sentential understanding and reasoning capabilities of language models, they fall short of detecting the precise nature of reasoning, i.e. semantic vs. pragmatic, due to the coarseness of their labeling schemas. In particular, these datasets conflate various implicational relations such as *entailment, implicature,* and *presupposition* under the same label i.e. *entailment*. However, linguists have long observed that entailments differ from implicatures and presuppositions specifically in terms of what kind of reasoning mechanisms underlie such implicational relations (Grice, 1975; Horn, 2006, 1972; Levinson, 2000; Sauerland, 2012).

A key distinction between entailments and implicatures is that of reasoning over *what is said* and *what is not said*. Entailment relations are inferences based on *what is said* and they arise as a consequence of the meanings of expressions in a sentence and the general laws of logic. The defining characteristic of an entailment relation between a premise (P) and a hyopthesis (H) is Truth. P entails H if and only iff whenever P is True H must be True as well. The P-H pair in (1) illustrates entailment. This is a logical corollary of the subset-superset relation between *fluffy cats* and *cats*.

(1) P *entails* H
P: Garfield is a fluffy cat.
H: Garfield is a cat.

Implicatures on the other hand are inferences based on what is not said and they follow from general cooperativeness principles of conversation (Grice, 1975, 1989). In (2), the relation between P and H is *implicature* but not *entailment.*

[1] https://github.com/kursathalat/ImplicaTR

Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024), pages 29–41

(2) P *implicates* H
Q: Is he handsome?
P: He is smart.
H: He is not handsome.

Unlike entailments, implicatures are not logical consequences of their premises. Instead, they arise through pragmatic reasoning. Implicatures can be distinguished from ordinary entailments by means of various tests such as *cancellation*, *suspension*, and *reinforcement*. For example, implicatures can be cancelled without leading to a contradiction but entailments cannot as illustrated in (3) and (4).

(3) **Entailment cancelled, contradiction**
P: Garfield is a fluffy cat.
H': Garfield is not a cat.

(4) **Implicature cancelled, no contradiction**
Q: Is he handsome?
P: He is smart... (H':) And handsome.

To test the pragmatic reasoning capabilities of language models in Turkish, we introduce ImplicaTR, the first fine-grained Turkish NLI dataset consisting of Premise-Hypothesis pairs containing *entailment, implicature, contradiction,* and *neutral* labels. We test various types of large language models (LLMs) using ImplicaTR and observe that LLMs are capable of carrying out both semantic and pragmatic reasoning with success rates of up to 98% accuracy. Despite their high levels of success, our ablation studies reveal that LLMs do not form a high level abstraction for pragmatic reasoning as they *cannot generalize* across various types of implicature contexts.

2 Related Work

NLI, a subset of the broader task known as Natural Language Reasoning (Yu et al., 2023), has been extensively researched within the context of textual entailment. Research in NLI led to the creation of numerous benchmark datasets aimed at training and evaluating the inferencing capabilities of language models. Major NLI datasets such as SNLI (Bowman et al., 2015) and (Williams et al., 2018) focused on three-way (entailment, contradiction, neutral) classification of inferential relations. Although these benchmark datasets have been widely adopted, they have also been noted to have some issues such as the predictability of the inference between premise and hypothesis due to repeating patterns within the hypothesis like negation (Gururangan et al., 2018; Poliak, 2020) or the overwhelming majority of upward entailing contexts leading the models to make errors in downward entailing contexts (Yanaka et al., 2019a). To overcome some of these challenges various NLI datasets have been created. (Yanaka et al., 2019b) created the HELP dataset to overcome the issues with downward entailment contexts. (Conneau et al., 2018) created the XNLI dataset to expand the NLI research into languages other than English. The availability of NLI datasets in Turkish is limited with NLI-TR (Budur et al., 2020), which presents an automatic translation of SNLI and MNLI combined, and with STSb-TR (Fikri et al., 2021) for semantic textual similarity.

Recent NLI research started to pay attention to more granular inference types that can help evaluate the precise reasoning capabilities of language models by distinguishing inference types such as *implicature, entailment, presupposition*. Implicature (George and Mamidi, 2020) and BIG-Bench (Srivastava and others, 2022) datasets were created for *particulatized implicatures*. Similarly, GRICE (Zheng et al., 2021) offers conversational reasoning and implicature data in the form of open dialogues devised by an automated grammar. The IMPPRESsive dataset (Jeretic et al., 2020) consists of semi-automatically generated *scalar implicatures* and *presuppositions* as Premise-Hypothesis pairs, where authors show that models can do pragmatic reasoning for some types of scales in their dataset.

This brief review of the literature reveals that the NLI literature needs more work in the areas of pragmatic reasoning and we aim to help fill this gap by investigating implicatures, which present a distinct line of work for the NLI research with its more granular comprehension of the pragmatic inferences. In addition, NLI research in Turkish has a limited scope, totally lacking an investigation into implicatures to the best of our knowledge. With its rich morphology and agglutinative nature especially reflected on the verbs, Turkish presents a peculiar case for probing into how implicatures are handled by NLI models.

3 Dataset: ImplicaTR

ImplicaTR is a semi-automatically generated Turkish NLI dataset annotated with a granular classification of sentential inference types covering *scalar implicatures* in addition to the conventional three-

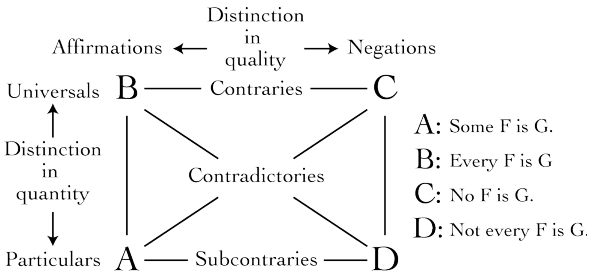

Figure 1: Square of Opposition

way NLI classes (*entailment, contradiction, neutral*). The dataset comprises five different linguistic categories (quantifiers, adjectives, verbs, modals, and numerals) with varying number of scalar pairs for each category.

3.1 Scalar Pairs

A scale (or a Horn Scale) (Horn, 1972) is a set of two or more lexemes that are in a relationship of strength or intensity. For instance, the scalar pair $\langle some, all \rangle$ contains the weaker term *some* and the stronger term *all*, between which there is a quantificational difference. Horn (2006) observed a set of logical relations between scalar elements (e.g. *some - all*) and their negations (*none - not all*) which he represented as quadruplets on a square of opposition shown in Figure 1.

In Figure 1, the universals B and C *entail* A and D, respectively, while B and C logically *contradict* each other. The particulars A and D are in a *neutral* relationship with their universal counterparts B and C. Notably, utterance of A or D *implicate* the truthfulness of one another. Thus, we obtain the conventional NLI classes along with the *implicature* inference from a quadruplet of sentences stemming from a scalar pair and their negation.

We created ImplicaTR by using a variety of scalar pairs and their negations as captured by the Square of Opposition. To ensure wide coverage, we covered a total of 44 scalar pairs from give distinct linguistics categories consisting of *adjectives, verbs, quantificational determiners, modal expressions*, and *numerals*. Some scalar pairs, as those in De Melo and Bansal (2013), were excluded as their scalar interpretations are highly contextual and impossible to control without further context.

3.2 Linguistic Categories

Scalar meanings in natural languages can be expressed by different lexical categories (e.g. adjectives, verbs, etc.) and yet the logical relations among scalar pairs are constant as noticed by linguists (Horn, 2006; Kennedy and McNally, 2005; Kennedy, 1999) and illustrated on the Square of Opposition in Figure 1. This indicates that humans are able to make abstract generalizations regarding the logical relations among scalar expressions regardless of their lexical categories or linguistic expression. To evaluate the abstract generalization capabilities of language models across different lexical categories, we used scalar pairs from five different categories: *adjectives*, *verbs numerals, modals* and *quantificational determiners*.

Adjectives and *verbs* form open-class categories. Open-class categories permit new members and cover a wider range of linguistic expressions compared to closed-class categories. Usually, this translates lower relative frequency per lexeme in a corpus compared to closed class categories. We used a total of 46 open-class words (28 adjectives and 18 verbs). Adjectival pairs include examples such as ⟨*benzer,aynı*⟩ ('similar-same'), ⟨*yakın, bitişik*⟩ ('close-adjacent'), whereas verbal pairs include instances such as ⟨*başla, bitir*⟩ ('start - finish') (following Jackendoff (1996); Pedersen (2014)).

Quantificational determiners, modals, and numerals form closed-class categories. Quantificational determiners are naturally scalar as they denote degrees of quantification. We used seven quantificational determiners to form various scalar pairs such as ⟨*birkaç, bütün*⟩ ('a few'- 'all'). Modal expressions also encode quantificational force (Hacquard, 2010) and thus create scalar pairs. Modal expressions come in various flavors such as *epistemic*, referring to the certainty of knowledge (Kaufmann et al., 2006), and *deontic*, referring to the cases of obligation or permission (Johanson, 2009). We have only used four epistemic modal expressions as deontic modals in Turkish usually result in ambiguity which makes it hard to evaluate the success of language models.

The last type of scalar expressions in the dataset are numerals. Numerals belong to closed-class words consisting of a finite number of lexical items yet they require particular attention for two key reasons. Numerals are by definition ordered and they form an infinite scale ($\langle 0, 1, 2, 3, 4, ... \rangle$ or ⟨*bir, iki, üç, dört, ...*⟩. This makes their distribution in any given dataset quite unbalanced. While some common numerals such as *bir, iki, beş, on* can be very frequent in a corpus, complex numeral expressions such as *üç yüz elli yedi* (357) or *on iki bin sekizyüz otuz üç* (12833) will be rare if present at

all. To alleviate the sparsity issue, we have limited the number of unique numerals in the dataset to 18 and we opted for relatively common numerals such as *bir, iki, beş, otuz, altmış, ...* (1,2,5,30,60,...) The second point to note is that numerals behave differently from other scalar expressions when they are combined with negation. In general, negation of a stronger value on a scalar pair implicates the weaker term. *"Not all chairs are dirty."* implicates *"Some chairs are dirty."* With numerals, negation of a stronger value raises two additional implicatures besides the implicature of the weaker value. These are *at-most* (Papafragou and Schwarz, 2005) and the *existential* implicatures as illustrated in (5).

(5) A: You need five apples for this dessert.
P: Oh, we don't have four apples.
H1: We have at most four apples.
H2: We have at least one apple.

3.3 Data Generation

ImplicaTR was built semi-automatically through an iterative process. For each scalar pair (<bazı, tüm> <some, all>), we manually created a few sample quadruplets of sentences ⟨A,B,C,D⟩, where sentence A contains the weaker term (bazı), B the stronger term (tüm), C negation of the weaker term (hiç), and D negation of the stronger term (tümü değil) as illustrated in Figure 2. A sample quadruplet is given in Table 1.

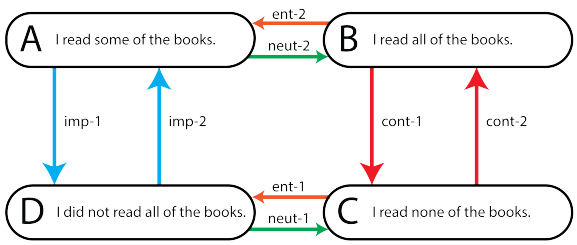

Figure 2: Quadruplets and inference relations

Table 1: A Sample Quadruplet

Sentence ID	Scalar Item	Sentence
A	bazı	Kitapların bazısını okudum.
B	tüm	Kitapların tümünü okudum.
C	hiç	Kitapların hiçbirini okumadım.
D	tümü değil	Kitapların tümünü okumadım.

In addition, we manually created a set of A sentences for each scalar pair that covers a wide range of linguistic structures. By using the manually created quadruplets as few-shot examples, we employed GPT-4 (OpenAI, 2024) to autogenerate the B, C, and D sentences for the remaining A sentences. At each iterative step, two expert linguists reviewed the autogenerated quadruplets to verify their grammaticality and the accuracy of the inference relations among the quadruplets. Scalar pairs that led to ambiguities and linguistic structures that disrupted the inference relations were removed after each iteration until we reached a reliable set of scalar pairs and linguistic structures. See Table 2 for a complete set of inferences obtained from a quadruplet.

In the final iteration, we created 19,350 sentence pairs covering the four types of inference types *entailment, implicature, contradiction* and *neutral*. The quality of the dataset was verified by randomly sampling 2,137 sentence pairs, ensuring a 95% confidence interval and a 2% margin of error. An expert linguist reviewed these sentence pairs, revealing that 97.89% of the data had correct inference labels. See Appendix A for the distribution of scalar pairs and other descriptive statistics about the dataset.

4 Experiment 1

Experiment 1 aims to explore whether LLMs exhibit pragmatic reasoning, specifically in scalar implicature resolution. We fine-tuned a series of models on ImplicaTR and observed that language models can successfully identify implicatures.

4.1 Experimental Setup

4.2 Data

We split the dataset into train (12,309 items), validation (3,153 items), and test (3,888 items) sets via stratified sampling to ensure that the model can see examples from each category and scalar pair and that a single quadruplet is included in only and only one of the splits.

4.3 Models

For this experiment, we used two different sets of models: Masked Language Models (e.g. BERT-family models) and generative models. BERT (Devlin et al., 2019) is an encoder-decoder model based on the transformers architecture (Vaswani et al., 2017). With their bidirectional architecture, BERT-family models take into account the left and the right context of a masked element within a sentence. On the other hand, generative LLMs based on transformers are trained on seq2seq tasks, where they take the input sequence and generate an out-

Table 2: A Sample Set of Inferences out of a Quadruplet

Premise Type	Hypothesis Type	Premise Example	Hypothesis Example	Inference Type/Label
A	D	Kitapların bazısını okudum.	Kitapların tümünü okumadım.	implicature
D	A	Kitapların tümünü okumadım.	Kitapların bazısını okudum.	implicature
C	D	Kitapların hiçbirini okumadım.	Kitapların tümünü okumadım.	entailment
B	A	Kitapların tümünü okudum.	Kitapların bazısını okudum.	entailment
D	C	Kitapların tümünü okumadım.	Kitapların hiçbirini okumadım.	neutral
A	B	Kitapların bazısını okudum.	Kitapların tümünü okudum.	neutral
B	C	Kitapların tümünü okudum.	Kitapların hiçbirini okumadım.	contradiction
C	B	Kitapların hiçbirini okumadım.	Kitapların tümünü okudum.	contradiction

put sequence; thus, these models learn and generate output by performing next-word prediction. We selected these two types of models as BERTs have been shown to demonstrate superior comprehension of language (Cho et al., 2021), while generative models are in widespread use in spite of their relatively poorer grasp of the linguistic insights (Fu et al., 2023; Raffel et al., 2023).

BERT-family models employed in this experiment are bert-base-uncased, BERT-NLI (Laurer et al., 2023), and BERTurk (Schweter, 2020). BERT-NLI is the DeBERTaV3-based zero-shot model and was trained on XNLI and MNLI datasets, which we expect would show greater performance on NLI tasks. BERTurk is a Turkish model and was trained on Turkish Wikipedia dumps, which allows us to compare the cross-task ability of this model against the cross-lingual ability of BERT-NLI. As for generative models, we fine-tuned the 7B parameter versions of Llama-2 (Touvron et al., 2023), Gemma (Team et al., 2024), and Mistral (Jiang et al., 2023). Training was done via prompting for generative models, for which a sample training item is given in Appendix B.

The training hyperparameters used for the BERT models are as given below.

Table 3: Training Hyperparameters for BERT Models

Hyperparameter	Value
hidden dropout value	0.3
attention dropout prob	0.25
number of epochs	10
gradient accumulation steps	2
warmup ratio	0.01
batch size	64
weight decay	0.05
learning rate	0.00001
lr reduction factor	0.5
lr reduction threshold	0.2

4.4 Results

We fine tuned the models on the training datasets and evaluated their success on the test sets. Figure 3 presents the accuracy scores of the fine-tuned models as well as the base models (before fine tuning). We observe that the base models are not biased towards any of the inference classes.

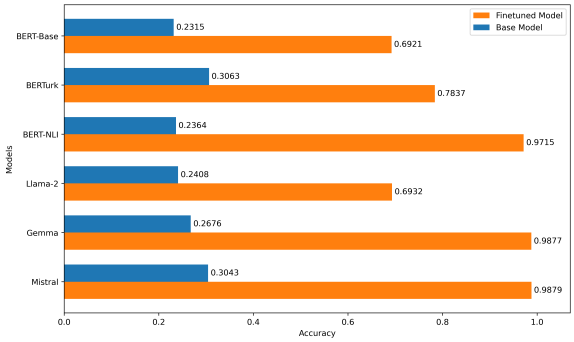

Figure 3: Accuracy scores from Experiment 1

Within the BERT family, BERT-NLI excelled the task with 0.97 while BERTurk achieves a higher score than the base model. This shows that the NLI training of BERT-NLI increased the ability of the model to recognize textual entailment even though we introduced a new class, *implicature*. Generative models demonstrated parallel results, where Gemma and Mistral reached accuracy scores of 0.98. These results suggest that generative models can handle pragmatic reasoning tasks such as detecting scalar implicatures. Llama-2 showed a poorer performance with 0.69, which we think is due to the the size the of the training data. Llama-2 was trained on 2T tokens whereas this number is 6T for Gemma and probably a similarly high number for Mistral. Therefore, models seem to learn the pragmatic contributions of words when they are exposed to them more during training.

4.5 Benchmark on XNLI and MNLI

In order to evaluate the performance of our fine-tuned models, we tested our fine-tuned BERT-NLI model on the XNLI and MNLI test sets as it was the best performing BERT model in our experiments. The original BERT-NLI model as well as XNLI and MNLI offers a three-way classification whereas our fine-tuned BERT-NLI model does more granular classification by predicting *implicature* as well. Thus, to evaluate the performance, we employed four different strategies in mapping our 4-way classification onto the 3-way classes of XNLI and MNLI test sets. First, without any alteration, we calculated the accuracy score by comparing predictions against ground labels as is. Then, we converted the implicature predictions to entailment, neutral and contradiction, and we calculated the accuracy score accordingly to see how the accuracy scores change per label.

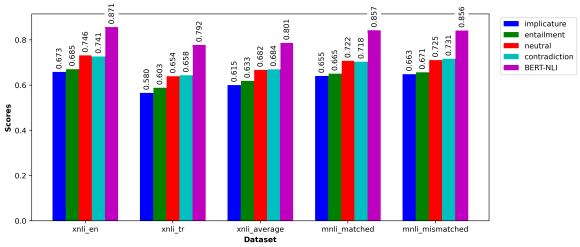

Figure 4: Accuracies of Finetuned BERT-NLI with Different Labeling on selected XNLI and MNLI sets, and Original BERT-NLI Score

The original *implicature* case has the lowest score in all sets while converting the predicted label to *neutral* or *contradiction* yielded the best scores. This is in line with the argument that NLI models are positively biased towards the *contradiction* and *neutral* classes because of the existence of negation words like *not* and superlative expressions denoting the maximal values (Gururangan et al., 2018). Compared to the base BERT-NLI model, our model's accuracy score is lower by 12%, which is expected due to the granularity of our labels and the smaller size of the implicature data.

5 Experiment 2

Upon observing that LLMs are capable of learning pragmatic inferences in the form of scalar implicatures, we conducted a second experiment where we perform an ablation study to test the generalization abilities of LLMs with respect to pragmatic reasoning in a supervised fashion. This experiment consists of two phases. In the first phase, we train five models by eliminating one of the linguistic categories entirely from training split in each model training and then test the model on the eliminated linguistic category. The goal is to test whether LLMs can create a sufficiently abstract generalization of scalar implicatures that can be used independent of the linguistic structures. In the second phase, we develop a sixth model by eliminating some scalar pairs from each linguistic category and test the model on the eliminated pairs. The goal in this second phase is to test the generalization abilities of LLMs within each category. The ablation study is followed by a feature analysis to inspect which linguistic features are influential in LLM performance in textual entailment and implicature reasoning.

5.1 Data Preparation

For the ablation study, we created five different splits. Table 4 presents the training and test categories for each model. We used stratified sampling to create training and validation splits to ensure that the model does not encounter any particular sentence in more than one split.

Table 4: Linguistic Categories Used for Training and Testing for Each Model

	Train	Test
Model-NUM	Adjectives Verbs Quantifiers Modals	Numerals
Model-MOD	Adjectives Verbs Quantifiers Numerals	Modals
Model-QUA	Adjectives Verbs Modals Numerals	Quantifiers
Model-VER	Adjectives Quantifiers Modals Numerals	Verbs
Model-ADJ	Verbs Quantifiers Modals Numerals	Adjectives

The second phase of the experiment involves MODEL-ALL, where some of the scalar pairs from

Table 5: Split sizes of models in Experiment 2

	Each Model in Phase 1	MODEL-ALL
	N of pairs	N of pairs
Train	11520	10560
Validation	2880	2640
Test	3600	4800
Total	**18000**	**18000**

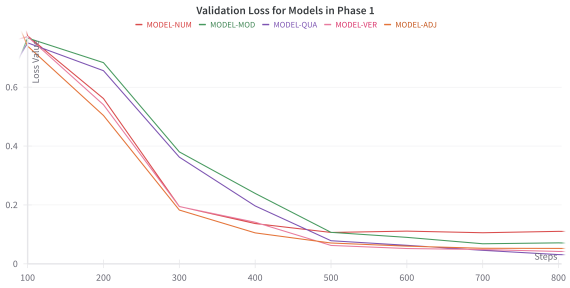

Figure 5: Validation Loss of Models in Phase 1

each category are removed from the training set (except for modals since the total number of pairs is very low). For example, all 30 quadruplets of <harmful, lethal> are left out in training and they are included in the test set of MODEL-ALL with a view to test whether the model can generalize what it learns for a specific linguistic category to the unseen scalar pairs within the same category. Data split sizes for all models are given in Table 5.

5.2 Ablation Study

We conducted Experiment 2 with the BERT-NLI model as it achieved the best performance in the Experiment 1 among the BERT models. Losses on validation set are plotted in Figure 5. While the elimination of *adjectives*, *verbs*, or *numerals* from the training data exhibits similar decrease patterns in loss, MODEL-MOD and MODEL-QUA values indicate that the absence of *modals* or *quantificational determiners* introduce a slight challenge for the model to learn the patterns but the models converge eventually.

Table 6: Chi-Square and Cramer's V Results

Pearson Chi-Square	p-value	Cramer's V
1256.2951	<0.0001	0.1525

5.3 Results

We evaluated each model on their respective test datasets and conducted a chi-square test to determine whether differences between categories are significant or not. The results indicated a significant differences between category results with a p-value <0.0001. Table 6 presents the results of the chi-square test and Table 7 reports the test scores for each model.

The results indicate that the models can successfully generalize to the categories of *modals* and *quantificational determiners* while we see moderate accuracy scores for *verbs* and *adjectives* and relatively low scores for *numerals*. We believe that these results are due to the distribution of scalar items in the pre-training data. *Modals* and *quantificational determiners are* closed-class expressions with relatively lower type frequencies (and thus higher token frequencies for each type). On the other hand, *adjectives* and *verbs* are members of open-class categories with relatively higher type frequencies (and thus lower token frequencies for each type). Finally, numerals have the largest type frequencies (theoretically infinite) despite being members of a closed-class category. Thus, the number of scalar relationships that a particular numeral can establish is also large (theoretically infinite), majority of which are unknown to the model or not reinforced in pre-training, which possibly decreases the model performance for numerals. These results suggest that the token frequency of a lexical item in the pre-training data is an important factor in a model's ability to execute pragmatic reasoning over expressions involving that lexical item. The results suggest that the tested LLMs may lack the ability to create sufficiently abstract generalizations for pragmatic reasoning that transcend particular linguistic structures.

In the second phase, we trained and evaluated MODEL-ALL in order to test the performance of the fine tuned NLI model on unseen scalar pairs within a previously trained category. The results are presented in Table 7.

MODEL-ALL suggests that the scalar reasoning exists within the linguistic categories for *adjectives* and *numerals*. Training on similar structures helped the model gain pragmatic reasoning capabilities to identify implicatures. Quantificational determiners also showed similarly accuracy scores. However, the model did not achieve high scores within the category of *verbs*. We believe that this

Table 7: Test results of models in Phase 1 and of respective linguistic categories in MODEL-ALL, where MODEL-ALL Accuracy scores specifically refer to the accuracy score of the linguistic category tested in the respective model from Phase 1. No score for modals as they are not tested in MODEL-ALL.

Model	Test Loss	Accuracy	F1	Precision	Recall	MODEL-ALL Accuracy
MODEL-NUM	1.5592	0.6036	0.5506	0.6723	0.6036	0.8541
MODEL-MOD	0.1416	0.9622	0.9621	0.9644	0.9622	-
MODEL-QUA	0.286	0.9336	0.934	0.9396	0.9336	0.9866
MODEL-VER	0.7137	0.7969	0.7948	0.8153	0.7969	0.6625
MODEL-ADJ	1.3374	0.7152	0.715	0.7169	0.7152	0.9733

might be due to the agglutinating nature of Turkish verbs (verbs usually occur with various suffixes on them) leading to a sparsity in the training data and impeding its generalization abilities.

5.4 Featural Significance Analysis

We followed up the ablation study with a featural significance analysis in order to unveil the potential linguistic triggers in our dataset that lead to the correct or incorrect classification of the premise-hypothesis pairs. For this, we first extracted a set of linguistic features and then fit logistic regression and random forest models to measure their impact on model performance.

In the NLI literature, various linguistic features have been argued to affect the model performance (Miaschi et al., 2020; Kriz et al., 2015; Talman et al., 2021; Wendland et al., 2021). Accordingly, we have included various features such as counts and lengths of certain tokens, predicate type, polarity, the word similarity within sentence, the similarity between premise and hypothesis, TF-IDF scores of the scalar items, the position of scalar item, and NER tags and sentiments in our analysis. The full list of features extracted is given in Appendix C . In a preliminary regression test, we observed that the NER and sentiment features had no impact on model performance; therefore, we excluded them from further analysis.

5.5 Logistic Regression

We fit a linear regression model with predictors as our extracted features and the outcomes as the prediction accuracy of the model. The linear regression model achieved an accuracy score of 0.80, which, we believe, makes the model appropriate for featural significance analysis. Figure 6 below presents the features with the most effect along with their coefficient scores.

The results suggest that the similarity between

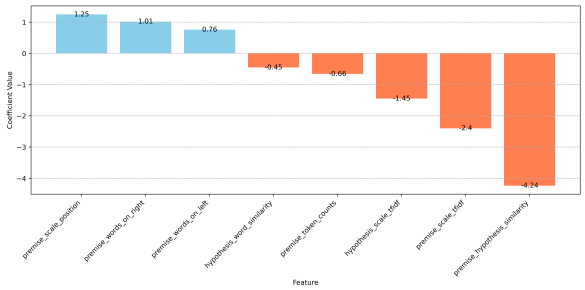

Figure 6: Feature coefficients of logistic regression model

a premise and a hypothesis and the high TF-IDF score of the scalar item in the premise sentence lowered the model performance. The feature 'premise_scale_position' refers to the position of the scalar item in the sentence. Given that Turkish is an SOV language and our dataset does not contain any word order inversions, we observe that closeness of the scalar item to the main verb improves the accuracy of the model. Although it goes beyond the scope of our current study to explain this observation properly, we speculate that this might be due to the pre-verbal position in Turkish being associated with new information focus (Göksel and Özsoy, 2000). In general, this position is reserved for new information in Turkish and new information is usually more attended to by speech participants. If LLMs are capable of associating the pre-verbal position with new information focus, they might be paying more attention to the scalar items in this position, leading to an increased accuracy.

5.6 Random Forest Model

We also fit a random forest model to further verify the effects of the features on the model prediction accuracy. For this model, we eliminated the features with low effect size and only used the continuous variables as predictors. The random forest

model achieved 0.79 accuracy and the coefficient results are in Figure 7.

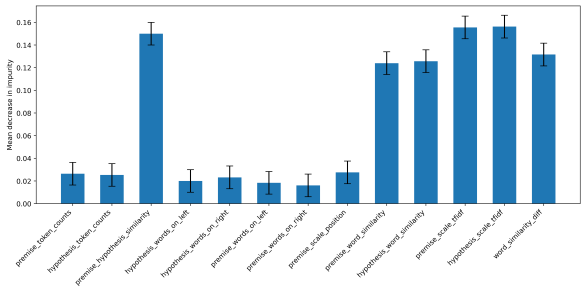

Figure 7: Feature importance of random forest model

We see that the results of the random forest model are in line with the regression analysis we did. In this model, where the coefficients are calculated by the decrease they cause in the mean accuracy (MDI), the features with the highest decrease are TF-IDF scores of the scalar items within the sentence. This is valid for both premise and hypothesis sentences, as the values of both are the highest. The similarities of the embeddings of the premise-hypothesis pair can be seen to have a negative effect on the correctness of the model prediction. Additionally, the average similarity scores of the words within a sentence are again one of the factors that decrease the score.

6 Conclusion

We presented ImplicaTR, a diagnostic dataset to test the pragmatic reasoning abilities of language models. ImplicaTR contains NLI-style sentence pairs with four distinct inference types, *entailment, contradiction, neutral* and *implicature*. We evaluated various LLMs and showed that they are capable of doing pragmatic reasoning and distinguishing *entailments* from *implicatures* with a high degree of accuracy. Our results also indicated that the models we tested cannot make sufficiently abstract generalizations across various linguistic structures for pragmatic reasoning and the type frequency of the scalar items is inversely correlated with the model success.

7 Limitations

This study introduces ImplicaTR and conducts two experiments on it to investigate the pragmatic capabilities of LLMs, but it also comes with a couple of limitations. First, while ImplicaTR is a diagnosis dataset, it is not a large one considering that it introduces a new class. Second, the genre and style of the items are not versatile, which might hinder the generalization capabilities of models. While the linguistic inquiry in Experiment 2 offers an insight into how models execute reasoning over implicatures, the features extracted can be extended to account for other syntactic and semantic phenomena.

8 Ethical Considerations

All sentence pairs used in ImplicaTR were generated synthetically, and no personal or sensitive information was used in order to ensure compliance with privacy standards and data protection regulations. Besides, efforts were made to minimize bias in the dataset by including a diverse range of linguistic expressions and contexts. We have made all code, models, and the dataset publicly available to promote transparency and reproducibility.

Acknowledgments

We thank Ömer Demirok, Cem Bozşahin, and an anonymous SIGTURK reviewer for their insightful comments at various stages of this work.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Emrah Budur, Rıza Özçelik, Tunga Gungor, and Christopher Potts. 2020. Data and Representation for Turkish Natural Language Inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8253–8267, Online. Association for Computational Linguistics.

Won Ik Cho, Emmanuele Chersoni, Yu-Yin Hsu, and Chu-Ren Huang. 2021. Modeling the Influence of Verb Aspect on the Activation of Typical Event Locations with BERT. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2922–2929, Online. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. *arXiv preprint*. ArXiv:1809.05053 [cs].

Gerard De Melo and Mohit Bansal. 2013. Good, Great, Excellent: Global Inference of Semantic Intensities.

Transactions of the Association for Computational Linguistics, 1:279–290.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint*. ArXiv:1810.04805 [cs].

Figen Beken Fikri, Kemal Oflazer, and Berrin Yanıkoğlu. 2021. Turkish dataset for semantic textual similarity. In *2021 29th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4.

Zihao Fu, Wai Lam, Qian Yu, Anthony Man-Cho So, Shengding Hu, Zhiyuan Liu, and Nigel Collier. 2023. Decoder-Only or Encoder-Decoder? Interpreting Language Model as a Regularized Encoder-Decoder. *arXiv preprint*. ArXiv:2304.04052 [cs].

Elizabeth Jasmi George and Radhika Mamidi. 2020. Conversational implicatures in English dialogue: Annotated dataset. *Procedia Computer Science*, 171:2316–2323.

Aslı Göksel and A Sumru Özsoy. 2000. Is there a focus position in turkish. *Studies on Turkish and Turkic languages*, 107:119–228.

H. P. Grice. 1975. Logic and Conversation. In Donald Davidson and Gilbert Harman, editors, *The Logic of Grammar*, pages 64–75.

H. P. Grice. 1989. *Studies in the way of words*. Harvard University Press, Cambridge, Mass.

Reto Gubelmann, Ioannis Katis, Christina Niklaus, and Siegfried Handschuh. 2023. Capturing the Varieties of Natural Language Inference: A Systematic Survey of Existing Datasets and Two Novel Benchmarks. *Journal of Logic, Language and Information*, 33(1):21–48.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Valentine Hacquard. 2010. Modality. *Language*, 86(3):739–741.

Laurence R. Horn. 2006. Implicature. In Laurence R. Horn and Gregory Ward, editors, *The Handbook of Pragmatics*, 1 edition, pages 2–28. Wiley.

Laurence Robert Horn. 1972. *On the Semantic Properties of Logical Operators in English*. Ph.D. thesis, University of California, California.

Ray Jackendoff. 1996. The proper treatment of measuring out, telicity, and perhaps even quantification in english. *Natural Language and Linguistic Theory*, 14(2):305–354.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are Natural Language Inference Models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint*. ArXiv:2310.06825 [cs].

Lars Johanson. 2009. 15. Modals in Turkic. In Björn Hansen and Ferdinand De Haan, editors, *Modals in the Languages of Europe*, pages 487–510. Mouton de Gruyter.

Stefan Kaufmann, Cleo Condoravdi, and Valentina Harizanov. 2006. Formal approaches to modality. In William Frawley, Erin Eschenroede, Sarah Mills, and Thao Nguyen, editors, *The Expression of Modality*, pages 71–106. Mouton de Gruyter.

Christopher Kennedy. 1999. GRADABLE ADJECTIVES DENOTE MEASURE FUNCTIONS, NOT PARTIAL FUNCTIONS.

Christopher Kennedy and Louise McNally. 2005. Scale Structure, Degree Modification, and the Semantics of Gradable Predicates. *Language*, 81(2):345–381.

Vincent Kriz, Martin Holub, and Pavel Pecina. 2015. Feature Extraction for Native Language Identification Using Language Modeling.

Moritz Laurer, Wouter Van Atteveldt, Andreu Casas, and Kasper Welbers. 2023. Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI. *Political Analysis*, 32(1):84–100.

Stephen C. Levinson. 2000. *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. The MIT Press.

Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2020. Linguistic Profiling of a Neural Language Model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 745–756. ArXiv:2010.01869 [cs].

OpenAI. 2024. GPT-4 Technical Report. *arXiv preprint*. ArXiv:2303.08774 [cs].

Anna Papafragou and Naomi Schwarz. 2005. Most Wanted. *Language Acquisition*, 13(3):207–251. Publisher: Taylor & Francis, Ltd.

Walter A Pedersen. 2014. *Inchoative verbs and adverbial modification: Decompositional and scalar approaches*. Ph.D. thesis, McGill University, Montreal.

Adam Poliak. 2020. A Survey on Recognizing Textual Entailment as an NLP Evaluation. *arXiv preprint*. ArXiv:2010.03061 [cs].

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv preprint*. ArXiv:1910.10683 [cs, stat].

Uli Sauerland. 2012. The Computation of Scalar Implicatures: Pragmatic, Lexical or Grammatical? *Language and Linguistics Compass*, 6(1):36–49.

Stefan Schweter. 2020. BERTurk - BERT models for Turkish.

Aarohi Srivastava et al. 2022. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint*. ArXiv:2206.04615 [cs, stat].

Aarne Talman, Marianna Apidianaki, Stergios Chatzikyriakidis, and Jörg Tiedemann. 2021. NLI Data Sanity Check: Assessing the Effect of Data Corruption on Model Performance. *arXiv preprint*. ArXiv:2104.04751 [cs].

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, and et al. Bulanova. 2024. Gemma: Open Models Based on Gemini Research and Technology. *arXiv preprint*. ArXiv:2403.08295 [cs].

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and et al. Fuller. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint*. ArXiv:2307.09288 [cs].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv preprint*. ArXiv:1706.03762 [cs].

André Wendland, Marco Zenere, and Jörg Niemann. 2021. Introduction to Text Classification: Impact of Stemming and Comparing TF-IDF and Count Vectorization as Feature Extraction Technique. In Murat Yilmaz, Paul Clarke, Richard Messnarz, and Michael Reiner, editors, *Systems, Software and Services Process Improvement*, volume 1442, pages 289–300. Springer International Publishing, Cham. Series Title: Communications in Computer and Information Science.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019a. Can neural networks understand monotonicity reasoning? *arXiv preprint*. ArXiv:1906.06448 [cs].

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019b. HELP: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 250–255, Minneapolis, Minnesota. Association for Computational Linguistics.

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2023. Natural Language Reasoning, A Survey. *arXiv preprint*. ArXiv:2303.14725 [cs].

Zilong Zheng, Shuwen Qiu, Lifeng Fan, Yixin Zhu, and Song-Chun Zhu. 2021. GRICE: A Grammar-based Dataset for Recovering Implicature and Conversational rEasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2074–2085, Online. Association for Computational Linguistics.

A Appendix A: Data Distribution

Table 8: ImplicaTR: Data Distribution

	N of scales	N of distinct terms	N of quadruplets per scale	N of sentences per quadruplet	Total N of Quadruplets	Total Pairs
Adjectives	15	28	30	8	450	3600
Verbs	9	18	50	8	450	3600
Quantifiers	9	7	50	8	450	3600
Modals	2	4	225	8	450	3600
Numerals	9	18	50	11	450	4950
Total					2250	19350

B Appendix B: Prompt Example

Below is an instruction that describes a classification task. Give a label in your response that appropriately completes the request.
You will give only the label.
Instruction:
The labels are:
Labels: entailment, neutral, contradiction, implicature
The two sentences that you will classify are:
Sentences: A: *Yeni kullanmaya başladığı ilaçlar zararlı değil.* B: *Yeni kullanmaya başladığı ilaçlar ölümcül.* **Question:** What is the correct label that describes the relationship of B to A?
Response:
contradiction

C Appendix C: Features

Group	Description	Names of Variables
Counts and Lengths	The counts of nouns and tokens, and average length of each token per premise-hypothesis	premise_noun_counts hypothesis_noun_counts premise_token_counts hypothesis_token_counts avg_premise_token_length avg_hypothesis_token_length
Verb	Whether the predicate is nominal or verbal	premise_is_root_verb hypothesis_is_root_verb
Polarity and Negation	The polarity of the sentence as obtained from the morphological markers on the root for premise and hypothesis. Also, the possible combinations between premise-hypothesis	premise_polarity hypothesis_polarity isPol_PosPos isPol_PosNeg isPol_NegPos isPol_NegNeg
NER	The NER tags obtained from both premise and hypothesis	CARDINAL, GPE, PERCENT, ORG, NORP, LOC, MONEY, QUANTITY, DATE, TIME, PERSON, LANGUAGE, EVENT, WORK_OF_ART, FAC, TITLE, ORDINAL
Sentiment	The sentiment as predicted by zero-shot as one of positive, negative, or neutral	sentiment_premise_negative sentiment_premise_neutral sentiment_premise_positive sentiment_hypothesis_negative sentiment_hypothesis_neutral sentiment_hypothesis_positive
Word Similarity	The average word similarity for each premise and hypothesis obtained from fastText and the difference between the two	premise_word_similarity hypothesis_word_similarity word_similarity_diff
Sentence Similarity	The similarity score premise-hypothesis pair calculated by the embeddings	premise_hypothesis_similarity
TF-IDF	TF-IDF score of the scalar item in each sentence for premises and hypotheses	premise_scale_tfidf hypothesis_scale_tfidf
Scalar Position	The respective position of the scalar item within a sentence along with the number of tokens to left and to the right for both premise and hypothesis	premise_scale_position premise_words_on_right premise_words_on_left hypothesis_scale_position hypothesis_words_on_right hypothesis_words_on_left