How do we get there? Evaluating transformer neural networks as cognitive models for English past tense inflection

Xiaomeng Ma
The Graduate Center, CUNY
xma3@gradcenter.cuny.edu

Lingyu Gao
Toyota Technological Institute at Chicago
lygao@ttic.edu

Abstract

There is an ongoing debate on whether neural networks can grasp the quasi-regularities in languages like humans. In a typical quasi-regularity task, English past tense inflections, the neural network model has long been criticized that it learns only to generalize the *most frequent* pattern, but not the *regular* pattern, thus can not learn the abstract categories of regular and irregular and is dissimilar to human performance. In this work, we train a set of transformer models with different settings to examine their behavior on this task. The models achieved high accuracy on unseen regular verbs and some accuracy on unseen irregular verbs. The models' performance on the regulars is heavily affected by type frequency and ratio but not token frequency and ratio, and vice versa for the irregulars. The different behaviors on the regulars and irregulars suggest that the models have some degree of symbolic learning on the regularity of the verbs. In addition, the models are weakly correlated with human behavior on nonce verbs. Although the transformer model exhibits some level of learning on the abstract category of verb regularity, its performance does not fit human data well, suggesting that it might not be a good cognitive model. [1]

1 Introduction

Many aspects of language can be characterized as quasi-regular: the relationship between inputs and outputs is systematic but allow many exceptions. English past tense inflection exhibits such quasi-regularity that the regular verbs follow the '*-ed*' rule (*help* - *helped*) and the irregular forms consist of a variety of changes such as changing vowel (*sing* - *sang*). There has been heated debate about how people represent regular and irregular for the past 40 years. For the single-route approach, Rumelhart and McClelland (1986) described a feed-forward connectionist neural model that learned both regular and irregular forms of the English verbs' past tense without explicit symbolic rules. However, this model received fierce criticisms from the proponents of the dual-route model (e.g. Pinker and Prince, 1988; Marcus et al., 1992), who argue that the speakers first reason over the abstract categories (regular - irregular), and process the regulars through rule-applying mechanism (adding *-ed*) and process the irregulars via gradient analogical processes. In addition, Pinker and Prince (1988) highlighted many empirical inadequacies of the model and argued that these failures stemmed from 'central features of connectionist ideology' and would persist in any neural network model.

With the advancement of deep learning in NLP, there has been renewed interest in the English past tense debate with modern neural networks. Kirov and Cotterell (2018) revisited the past tense debate and showed that modern recurrent encoder-decoder (RNN) neural models overcame many of the criticisms. Their model achieved near-perfect accuracy on the unseen regular verbs and some accuracy on the unseen irregular verbs (28.6% as 5 correct irregular verbs). In addition, the model's results on the nonce verb inflections correlate with human experimental data (Spearman's $\rho = 0.48$ for regulars and $\rho = 0.45$ for irregulars). Thus they concluded that the neural model could be a cognitive model. However, other studies have shown that the modern neural network is still susceptible to the criticism raised by Marcus et al. (1995): the neural models lack symbolic rule learning ability and are vulnerable to the frequency distribution of the data, so they may learn to extend the *most frequent* pattern, instead of the *regular* pattern. Corkery et al. (2019) closely examined the model's performance on the nonce verbs and found that the fit to the human data is weak, especially for the irregular

[1] The code and data for this paper can be found at: https://github.com/xiaomeng-ma/English-Past-Tense

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1101–1114

verbs. Similarly, McCurdy et al. (2020) used German plural to demonstrate that the RNNs tend to overextend the most frequent plural class to nonce words and do not match the human speakers' data. Beser (2021) found that in English and German plurals, transformers are also susceptible to the frequency distribution of the data as RNNs. Prior work has generally focused on the comparison between model's performance and human behavior on nonce verbs, and few have explored the neural model's behavior on English regular and irregular verbs.

In our study, we closely examine the transformer's behavior on English past tense inflections corresponding to the training data's regular-irregular type and token frequency distributions to explore whether the models learn and apply symbolic rules. We train a set of transformers with different frequency distributions and experiment with resampling the training data for each epoch (§4). On our evaluation (§5.1) of English verbs, the transformers achieved over 95% accuracy on unseen regulars and some accuracy on unseen irregulars (ranging from 0% - 22%). We find that models exhibit different behaviors on the regulars and the irregulars, that the performance on regulars is more affected by the type frequency but not token frequency, and vice versa for the irregulars, suggesting that the models have some degree of abstract representation of verb regularity. We observe that the majority of the errors can be attributed to misclassification (e.g., treating an irregular as regular), with a smaller proportion of errors caused by applying the wrong inflection. For nonce verb evaluation (§5.3), the models vary in correlations with human data. Generally, the models correlate with human data better on regulars than irregulars, but the overall correlations are weak. In conclusion, we found that the transformer models display some degree of abstract representation of verb regularity, but do not fit human data well, thus can not be a good cognitive model.

2 Hypotheses and Predictions

2.1 Hypotheses

We aim to investigate the transformer's ability to generalize symbolic categories and rules in English past tense inflection task. Wei et al. (2021) proposed three hypotheses for how a neural network processes the symbolic rules by analyzing the behavior of BERT model (Devlin et al., 2019) on subject-verb agreement in English. We adapted their hypotheses and combined the theories of past tense debate to form our hypotheses. **H1: Idealized Symbolic Learner** operates over abstract categories and rules. For example, if x is a REGULAR verb and x ends with /d/ or /t/, then PAST(x) = x + /ɪd/. This is also the hypothesis for how humans process the regulars in the dual-route model. Under this hypothesis, the model would not misclassify verbs and is only sensitive to the type frequency, but not token frequency[2]. **H2: Naive Pattern-Associating Learner** does not necessarily represent any abstract features of the input verbs (such as regular/irregular); instead, it produces the output by a neuron-like activation process, which is analogous to an early feed-forward network as proposed in Rumelhart and McClelland (1986). This is the foundation for modern transformers, because transformer models also incorporate feed-forward layers. Therefore, the transformer model would naturally fall under this hypothesis. **H3: Symbolic Learner with Noisy Observations** is a hybrid of H1 and H2, suggesting that the model at its core is a symbolic learner, but with noisy observations. The model is able to generalize the abstract category for regular and irregular verbs, as well as the inflection patterns. However, the noisy observations would affect its ability to map the inputs to the correct category and/or apply the appropriate past tense inflection. Under this hypothesis, the model's categorization ability is mainly affected by the type frequency, and the pattern generalization is affected by both type and token frequency.

In this work, we expect the transformers to behave like H3, which operates based on pattern-associating and shows some level of symbolic learning. Moreover, the behavior on regular verbs should be a **STRONG Symbolic Learner with LESS noisy observations**, since the majority of English verbs are regular verbs and the regular inflection (adding /-d/, /-t/ or /ɪd/) can be easily summarized as a rule. The behavior on the irregular verbs should be a **WEAK Symbolic Learner with MORE noisy observations**, given that there are less than 200 irregular verbs in English with many implicit irregular inflection patterns (e.g., *go-went*).

[2]Wei et al. (2021) suggested that the 'idealized symbolic leaner would not be affected by word specific properties such as frequency', which we interpret as token frequency. In addition, psycholinguistic studies also suggested that human learners generalize phonological patterns based on type frequency and ignore the token frequency (e.g. Bybee, 2003)

2.2 Predictions and Summary of Findings

Since H2 is the basis of transformer models, we need to show that the model shows some symbolic learning ability to confirm H3. Evidence for symbolic learning includes type frequency effects and accurately classifying verbs into regulars and irregulars. In addition, we also need to demonstrate that the models exhibit stronger symbolic learning ability on regulars than irregulars. We would expect the regulars to display a strong type frequency effect and a weak token frequency effect, and vice versa for the irregulars. In addition, H3 learner predicts that the errors are due to failures to identify the verb as a regular verb, and/or apply the appropriate inflection.

Our experiments (§5.1) show that both regular and irregular verbs exhibit a clear type frequency effect and the models achieved good classification accuracy, suggesting some degree of symbolic learning. In addition, the regulars are more affected by the type frequency but not token frequency (and vice versa for the irregulars), suggesting that the regulars demonstrate stronger symbolic learning ability than the irregulars. The analysis also found misclassification errors and wrong inflection errors for the regulars and irregulars.

3 Data

The base dataset is the same one used in previous studies with English past tense, which includes 4,039 English verbs from the CELEX database (Baayen et al., 1995). We converted the verbs to IPA symbols based on Carnegie Mellon University Pronouncing Dictionary using `eng-to-ipa` python package,[3] and checked each verb's past tense forms on Merriam Webster dictionary.[4] Among these verbs, 3,857 are regular verbs; 150 are irregular verbs; and 32 verbs have both regular and irregular forms, e.g. *knit* - *knit* or *knitted*.[5] We also created two labels for each verb: Regularity and Verb class. The regularity indicates whether the verb is regular or irregular. The verb class corresponds to the inflection of each verb, which includes three classes for regular verbs (/-d/, /-t/, /-ɪd/) and seven classes for irregular verbs, including vowel change, vowel change +/-d/, vowel change +/-t/, ruckumlaut, weak, level and other (Cuskley et al., 2015).

	Example	Count	%
Regular		3857	95.5
/-d/	called	2045	50.6
/-t/	worked	763	18.9
/-ɪd/	wanted	1049	26.0
Irregular		182	4.5
vc	hide-hid	125	3.1
vc+/-t/	feel-felt	12	0.3
vc+/-d/	tell-told	10	0.2
ruck	buy-bought	8	0.2
weak	send-sent	9	0.2
level	quit-quit	11	0.3
other	go-went	7	0.2

vc = vowel change, ruck = ruckumlaut

Table 1: The regularity and verb class distribution in the CELEX dataset (the ambiguous verbs are treated as irregulars).

The examples for verbs of different regularities and verb class labels in the base dataset are shown in Table 1.[6]

3.1 Test Data

We evaluated the models on two test datasets: nonce verbs and real English verbs. Following the previous studies, we used 58 nonce verbs in Albright and Hayes (2003) for comparison with human behavior. For the real English verb test dataset, we randomly selected 80 verbs from the CELEX database, including 60 regular verbs (20 per verb class) and 20 irregular verbs (2 verbs from vowel change + /-t/ class and 3 verbs from other classes).

3.2 Training Data

After excluding the verbs in the test data, we developed 4 training datasets based on type frequency and token frequency. In the type frequency based training datasets, each verb appears only once. Since there are 32 ambiguous verbs, we create TYPE$_{reg}$ where these verbs are all treated as regular, and TYPE$_{irr}$ where they are all treated as irregular.

Then we created TOKEN$_{both}$, a token frequency based dataset with each verb appearing based on its CELEX frequency, where 'both' indicates that we consider both regular and irregular forms for ambiguous words. For example, the irregular form *knit* appears 5 times, and regular form *knitted* appears 12 times. As regular verbs dominate all these 3 datasets, we created TOKEN$_{irr}$, where only the irregular verbs appear based on their CELEX fre-

[3] https://pypi.org/project/eng-to-ipa/

[4] https://www.merriam-webster.com/

[5] The counts are different from Kirov and Cotterell (2018) because the original dataset has some inconsistent labeling. Details are explained in Appendix.

[6] The 32 ambiguous verbs are treated as irregular in the table.

Training set		Regular	Irregular	Total tokens
Type based	TYPE$_{reg}$	96.6%	3.4%	3,959
	TYPE$_{irr}$	95.9%	4.1%	3,959
Token based	TOKEN$_{both}$	68.7%	31.3%	147,711
	TOKEN$_{irr}$	7.7%	92.3%	49,983

Table 2: Regular and irregular verb distribution in different training datasets.

Input	Start, k, ɔ, l, End
Model	**Output**
BASE	Start, k, ɔ, l, d, End
LABEL$_{reg}$	Start, reg, k, ɔ, l, d, End
LABEL$_{vc}$	Start, +d, k, ɔ, l, d, End
LABEL$_2$	Start, reg, +d, k, ɔ, l, d, End

Table 3: Input and gold output in the training data with different labels for the verb 'call', tokens are separated by comma.

Resample	Count$_{Reg}$	Count$_{Irr}$	Irr. ratio (%)
BALANCE	129	129	50.0
REG$_{ds}$	48	129	72.6
IRREG$_{ds}$	283	129	31.3

Table 4: Count of regular (Reg) and irregular (Irr) verbs in three epoch training datasets. Irr. ratio denotes the percentage of irregular verbs in training data per epoch.

quency, and the regular verbs all appear once, of which the irregular rate is 92.3%. The regular and irregular rates for all training sets are shown in Table 2.

4 Experiment

4.1 Transformer Models

We used the sequence-to-sequence transformers (Vaswani et al., 2017) to generate the past tense of the root verbs trained from scratch. Our BASE model used the IPA phonemes of the root verb to generate the past tense inflections. We further examined whether identifying the regularity and verb class before generating the past tense would improve the model's performance. We added LABEL$_{reg}$ for regularity, LABEL$_{vc}$ for verb class, and LABEL$_2$ for both. Examples of input and gold output in the training data are shown in Table 3.

Since there are less than 200 irregular verbs in English, the model will be inevitably biased towards the regulars on type-based datasets. To adjust this imbalanced distribution, we downsample the number of regular verbs to match the number of irregulars in training data per epoch on TYPE$_{irr}$, which we called BALANCE.[7] To investigate the type-frequency effect, we further apply two unbalanced resampling methods per epoch:[8] REG$_{ds}$ downsizes the regulars to match the decreased regular rate in Parents' Data.[9] , and IRREG$_{ds}$ downsizes the irregulars to match the irregular rate in TOKEN$_{irr}$. Count of regular and irregular verbs, as well as irregular ratio seen per training epoch are listed in Table 4.

In addition, we added a pointer-generator mechanism (Vinyals et al., 2015) to the transformer model to reduce bizarre errors like **membled* for *mailed* that was reported in Rumelhart and McClelland (1986)'s original model[10]. This model could choose between generating a new element and copying an element from the input directly to the output. Transformers with copy mechanism have been used for word-level tasks (Zhao et al., 2019) and character-level inflections (Singer and Kann, 2020).

4.2 Experiment Setups

Both encoder and decoder of our models have 2 layers, 4 attention heads, 128 expected features in the input, and 512 as the dimension of the feed-forward network model. For training, we split the dataset into train-dev splits of 90-10, set model dropout to 0.1, and used Adam optimizer (Kingma and Ba, 2014) with varied learning rate in the training process computed according to Vaswani et al. (2017). Besides, we set batch size to 32 for type-based datasets, 64 for TOKEN$_{irr}$, and 128 for TOKEN$_{both}$. We run 30 epochs for all datasets. When we apply resampling methods (BALANCE, REG$_{ds}$, and IRREG$_{ds}$), we set batch size to 8 and run 100 epochs,

[7]There are 162 irregular verbs (excluding 20 verbs in test) in TYPE$_{irr}$. The train-dev split is 80-20, yielding 129 irregular verbs in training. We choose TYPE$_{irr}$ as it contains the most number of unique irregulars.

[8]We keep the numbers of irregular verbs unchanged, as we would prefer the model to see all irregular verbs for higher accuracy on irregulars.

[9]We selected 8 children's corpora in the CHILDES database (MacWhinney, 2000) and aggregated their parents' past tense verbs. If we leverage the percentage of its irregulars with the same construction method of TOKEN$_{irr}$, the irregular rate is 72.6%. Details are shown in Appendix 7.1.1.

[10]Kirov and Cotterell (2018) also reported one instance of this type of error and suggested that this type of errors could be eliminated by increasing training epochs. This type of errors has also been reported in other inflection tasks such as text normalization (Zhang et al., 2019).

Train Set	Model	Regular van.	Regular copy	Irregular van.	Irregular copy
TYPE$_{reg}$	BASE	99.0	99.0	**4.0**	0.0
	LABEL$_{reg}$	97.3	**99.7**	0.0	**1.0**
	LABEL$_{vc}$	**99.3**	98.3	1.0	**1.0**
	LABEL$_2$	99.0	**99.7**	1.0	0.0
TYPE$_{irr}$	BASE	97.0	97.0	**2.0**	**3.0**
	LABEL$_{reg}$	**99.0**	**99.7**	0.0	1.0
	LABEL$_{vc}$	94.7	99.3	0.0	0.0
	LABEL$_2$	97.0	97.7	0.0	1.0
TOKEN$_{both}$	BASE	**98.0**	**99.3**	**11.0**	**8.0**
	LABEL$_{reg}$	96.7	97.0	10.0	4.0
	LABEL$_{vc}$	97.7	97.0	2.0	2.0
	LABEL$_2$	**98.0**	97.0	4.0	3.0
TOKEN$_{irr}$	BASE	**95.7**	96.0	**22.0**	4.0
	LABEL$_{reg}$	95.0	**97.7**	9.0	**12.0**
	LABEL$_{vc}$	93.0	96.3	5.0	10.0
	LABEL$_2$	95.0	94.3	6.0	5.0

Table 5: Test accuracy (%) for our models for regular and irregular verbs, where 'van.' and 'copy' refer to the vanilla transformer model and the transformer model with pointer-generator mechanism respectively.

as there's fewer data per training epoch. As most of the datasets are highly unbalanced, we compute accuracy for both regular verbs and irregular verbs on dev set, and average them to select the best model. For inference, we set beam size to 5.

5 Results

5.1 English verbs' Test Accuracy

We calculated the test accuracy of our models based on the regulars and irregulars in the real English verb test set, which is shown in Table 5.[11] For all models, the regular verbs' accuracy was over 93%, and the irregular accuracy ranges from 0%-22% where the token-based models have better accuracy. The copy mechanism improved the accuracy for regular verbs, as we expected. The LABEL$_{reg}$, LABEL$_{vc}$, and LABEL$_2$ did not improve the irregulars accuracy for the vanilla model. The accuracy for each verb class can be found in Appendix Table 15 and Table 16.

Testing H1: Evidence for Symbolic Learning To show that the models exhibit some level of symbolic learning, we first examine the test accuracy of resampling method to explore the type frequency effect. As shown in Table 6, the accuracies of the regular verbs increase as their type frequency and ratio increase, showing the type frequency effect. In addition, the irregular verbs exhibit a relative type frequency effect too, that the accuracy increases as the type ratio increases, while the absolute frequency remains the same.

[11] All accuracy in this paper are averaged over 5 runs with different random seeds, while errors are counted by summing up the errors of different runs.

Test Acc	Model	Regular van.	Regular copy	Irregular van.	Irregular copy
BALANCE	BASE	72.7	**74.7**	23.0	**24.0**
irr:129	LABEL$_{reg}$	71.0	62.3	**24.0**	21.0
reg:129	LABEL$_{vc}$	68.7	71.3	17.0	18.0
	LABEL$_2$	**74.0**	68.7	19.0	14.0
REG$_{ds}$	BASE	**58.7**	**61.3**	**32.0**	25.0
irr:129	LABEL$_{reg}$	56.7	52.7	23.0	**28.0**
reg:48	LABEL$_{vc}$	56.0	52.0	21.0	20.0
	LABEL$_2$	55.7	60.3	21.0	15.0
IRREG$_{ds}$	BASE	77.0	**85.3**	**21.0**	15.0
irr:129	LABEL$_{reg}$	82.3	73.7	16.0	15.0
reg:283	LABEL$_{vc}$	**83.3**	72.7	14.0	**16.0**
	LABEL$_2$	79.7	81.7	12.0	10.0

Table 6: Test accuracy (%) for models trained on resampled data of TYPE$_{irr}$, where van. refers to vanilla model without copy mechanism. The irregular and regular tokens per epoch are listed for each resampling method.

Label Acc	Model	Regular van.	Regular copy	Irregular van.	Irregular copy
BALANCE	LABEL$_{reg}$	77.3	72.3	**79.0**	**85.0**
	LABEL$_2$	**85.3**	**83.7**	61.0	72.0
REG$_{ds}$	LABEL$_{reg}$	60.7	**72.0**	**90.0**	**88.0**
	LABEL$_2$	**66.0**	65.3	87.0	82.0
IRREG$_{ds}$	LABEL$_{reg}$	**90.0**	82.0	54.0	**59.0**
	LABEL$_2$	85.3	**87.7**	**55.0**	55.0

Table 7: Regularity label accuracy (%) for models with different resampled methods.

We further calculated the regularity label's accuracy on LABEL$_{reg}$ and LABEL$_2$ to examine the model's ability to categorize verbs into regulars and irregulars. As shown in Table 7, the models achieved good label accuracy for both regulars and irregulars, suggesting that the model has the ability to correctly classify the verbs. The label accuracies also display a type frequency effect, that the accuracies increased as the type frequency and ratio increased. These findings confirm that the model exhibits some level of symbolic learning.

Regular vs Irregular: Strong vs Weak Symbolic Learner We first examine the type and token frequency effect on the regulars and irregulars. The regular accuracy should be affected more by the type frequency than the token frequency, and vice versa for the irregulars. For the type frequency

Accuracy change		mean±std	max
Type Freq. Effect	reg	1.2±2.0	4.7
TYPE_{reg}- TYPE_{irr}	irreg	0.1±1.6	-3.0
Token Freq. Effect	reg	0.1±2.2	-3.0
TYPE_{irr}- TOKEN_{irr}	irreg	-4.6±3.2	-10.0

Table 8: The accuracy change (%) for type frequency effect comparison (TYPE_{reg}- TYPE_{irr}) and token frequency comparison (TYPE_{irr}- TYPE_{reg}).

effect, we calculated the accuracy change for different models of TYPE_{reg} and TYPE_{irr} in Table 5. The regular's accuracies are more affected by the change of type frequency than the irregulars, with higher average change and max change, as listed in Table 8. For token frequency effect, we calculated the accuracy change in TYPE_{irr} and TOKEN_{irr} where the regular and irregular's type frequency remains the same, but token frequency increased in both training datasets. The irregulars are more affected by the change of token frequency than the regulars, as listed in Table 8.

Next, we examine the model's classification ability. We manipulate the inferencing process for LABEL_{reg} and LABEL_2 models by manually setting the regularity label to the gold label[12] and let the model output the past tense based on the correct category. This method allows us to explore how classification affects test accuracy. The accuracy results for different models after inferencing is listed in Table 9. Inferencing improved the accuracy for the irregulars more than the regulars. This result indicates that misclassification errors are frequent for irregulars, but not regulars, suggesting that the models have a stronger classification ability for the regulars than the irregulars.

In summary, the transformers exhibit stronger symbolic learning ability on the regulars than the irregulars that regular accuracy is more affected by type frequency but not token frequency, and vice versa for the irregulars. The models made fewer errors due to classification on the regulars than the irregulars.

5.2 Error Analysis

We further conduct error analysis on regular and irregular verbs. H3 predicts the model to make classification errors as well as inflection pattern

Train Set	Model	Regular		Irregular	
		van	copy	van	copy
TYPE_{reg}	LABEL_{reg}	98.7	99.7#	22.0	14.0
	LABEL_2	99.0#	100.0	36.0	24.0
TYPE_{irr}	LABEL_{reg}	99.0#	99.7#	29.0	21.0
	LABEL_2	98.3	99.0	39.0	32.0
TOKEN_{both}	LABEL_{reg}	99.0	99.7	54.0	31.0
	LABEL_2	99.7	100.0	56.0	53.0
TOKEN_{irr}	LABEL_{reg}	98.3	100.0	50.0	30.0
	LABEL_2	99.0	99.3	48.0	57.0

Table 9: Test accuracy (%) after inferencing by setting the regularity label to the gold label. # indicates no change compared to the test accuracy without inferencing in Table 5.

Regular Error	Counts	Example
classification	144 (57.3%)	fine: /faʊn/
inflection	15 (6.0%)	coach: /koʊʧd/
copy	92 (36.7%)	unleash: /əniʃt/
Irregular Error	**Counts**	**Example**
classification	2755 (89.8%)	seek: /sikt/
inflection	279 (9.1%)	abide: /əbaʊd/
creative	34 (1.1%)	forgo: /fɔrgru/

Table 10: The counts and examples of regular error types and irregular error types. Counts are computed by summing up errors of all the models listed in Table 5.

errors. The regulars should have a lower percentage of both types of errors than the irregulars, since it is a STRONGER symbolic learner with less noisy observations.

We categorized the regular and irregular errors into classes based on the H3's prediction: **1. classification errors**, where the model output an irregular form for a regular verb, or a regular form for the irregular, **2. inflection errors** where the model applied a wrong regular inflection to a regular verb or a wrong irregular inflection to an irregular verb. In addition, for regular verbs, we also found **copy errors** where the model copied the verb root incorrectly, and **creative errors** for the irregulars where the model output some unseen inflection patterns. All errors of the models in Table 5 are manually annotated by researchers with linguistic training. The counts and examples for each error type are listed in Table 10. The proportions of classification and inflection errors are lower for the regulars than the irregulars, further providing evidence for regular as STRONG symbolic learner.

We further examined the copy errors for the regular verbs. Most of the errors either omit a conso-

[12]For example, for the verb *rethink*, the LABEL_{reg} will first output the 'reg' or 'irreg' label before producing the past tense. We manually set the label to 'irreg' and let the model predict based on the set label.

nant if two consonants are next to each other, e.g. *unleash*: /ənɪʃt/, *hitch**h**ike*: /hɪʧaɪkt/, or omitting a vowel if two vowels appear adjacent, e.g. *tri**u**mph*: /traɪmft/, *co-**o**pt*: /koʊptɪd/. This pattern suggests that the models might have learned that consonant or vowel clusters are not likely to appear in English, thus adjusting its output to avoid improbable consonant and vowel clusters.

5.3 Nonce verbs' correlation with humans

In this section, we compared the models' performance with human behavior by correlating the results on nonce verbs. The human experiment data is from two experiments run by Albright and Hayes (2003). They created 58 nonce English verbs and assigned regular and irregular past tense forms to each verb, e.g., *bize*: /baɪzd/, /boʊz/. 16 of these verbs were assigned 2 irregular forms, e.g., *rife*: /roʊf/ and /rɪf/. The participants were asked to first produce the past tense forms of these verbs, resulting in a production probability (P_{pro}), and to rate the regular and irregular forms of the past tense verbs, yielding a rating score. We follow Corkery et al. (2019)'s practice by treating each model as an individual participant and using the aggregated results to compare with the human results. To calculate the model's production probability, we used top-k sampling method to generate the top 5 outputs for each nonce verb, and aggregated the results over 5 random seeds. The model's production probability of each verb form is aggregated over 25 outputs. We correlated the model's P_{pro} with human's P_{pro} using Pearson r and used Spearman ρ to correlated the model's P_{pro} and humans' rating score.

The correlations with human data vary a lot among our models with different settings, i.e., some models could achieve a correlation over 0.7, while other models have negative correlations with human's data. The summary of the correlations' statistics of all the models is listed in Table 11. Detailed correlation for each model can be found in Table 17 in Appendix. The LABEL$_{vc}$ + TOKEN$_{both}$ model (vanilla LABEL$_{vc}$ trained on TOKEN$_{both}$) achieves the best overall correlation with human data, as is listed in Table 12. This model has a higher correlation with regular verbs than irregular verbs. For the models trained on resampled data, the BASE + BALANCE (vanilla BASE model with BALANCE resampling method) achieved the best overall correlation, as listed in Table 13.

		Mean	Std	Range
Regular	$P_{pro} r$	0.31	0.29	[-0.19, 0.70]
	Rate ρ	0.48	0.21	[0.02, 0.79]
Irregular	$P_{pro} r$	0.32	0.13	[0.06, 0.62]
	Rate ρ	0.31	0.12	[-0.06, 0.55]
Irregular 2	$P_{pro} r$	0.25	0.28	[-0.25, 0.77]
	Rate ρ	0.18	0.16	[-0.25, 0.61]

Table 11: The mean, standard deviation, and range for the correlation of different models (including all the models in Table 5 and the models in Table 6). Irregular 2 stands for the 16 verbs with 2 irregular forms. P_{pro} represents the production probability.

LABEL$_{vc}$ + TOKEN$_{both}$	P_{pro} (r)	Rating (ρ)
Regular (N = 58)	0.57	0.59
Irregular (N = 58)	0.22	0.22
Irregular 2 (N = 16)	0.12	0.36

Table 12: The correlations with human's data for vanilla LABEL$_{vc}$ trained on TOKEN$_{both}$.

BASE + BALANCE	P_{pro} (r)	Rating (ρ)
Regular (N = 58)	0.62	0.74
Irregular (N = 58)	0.44	0.45
Irregular 2 (N = 16)	0.69	0.28

Table 13: The correlations with human's data for vanilla BASE model with BALANCE resampling method.

In addition, we plotted LABEL$_{vc}$ + TOKEN$_{both}$, BASE + BALANCE and human's production probability for each nonce verb in Figure 1. Human speakers are generally able to produce some irregular forms for the nonce verbs, except for only one verb (*nace*). The models are less flexible in producing irregular forms. The LABEL$_{vc}$ + TOKEN$_{both}$ model only produced the regular forms for 27 verbs and 36 verbs for the BASE + BALANCE model. For the verbs with 2 irregular forms, humans are able to produce both forms for most of the verbs except for 3 verbs. However, the models' behaviors are more extreme that they are more likely to output only one type of irregular form of the verb. In addition, models and humans both produced many 'other' forms that are not included in Albright and Hayes (2003). For models, the 'other' forms are usually alternative irregular forms. For example, for the verb 'shee' /ʃi/, model's 'other' output include /ʃɛ/, /ʃɔ/, /ʃit/. Due to a lack of description of the 'other' output in human data, we could not closely examine whether model's other outputs are similar to humans.

In conclusion, it's difficult to make a simple state-

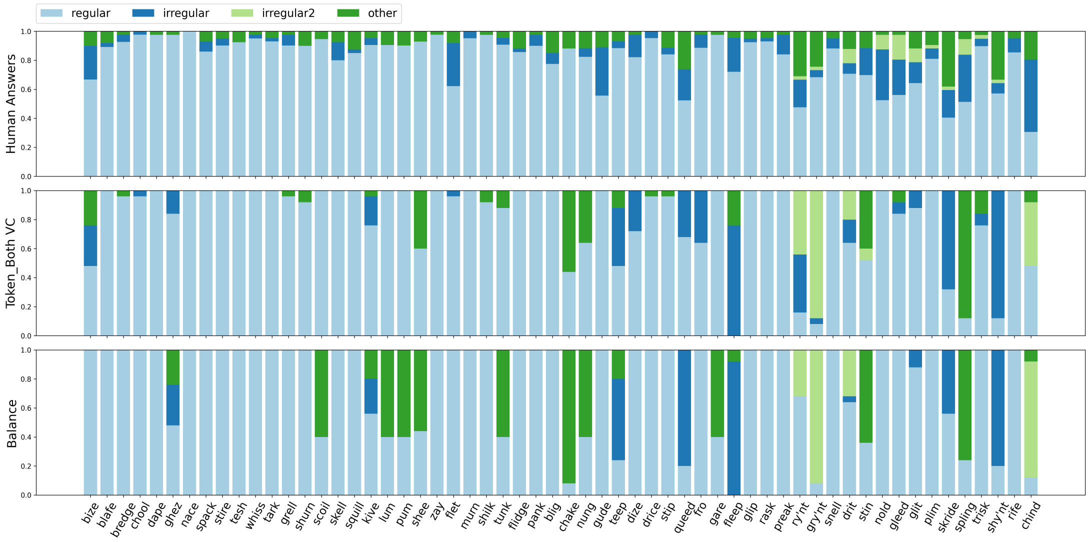

Figure 1: Percentage of regular, irregular, irregular 2 and other responses produced by humans (top), LABEL$_{vc}$ + TOKEN$_{both}$ model and BASE + BALANCE model. The last 16 verbs (starting with 'preak') have 2 irregular forms.

ment whether the model behaves like the human. Our best-performing models are able to achieve a high correlation with regular verbs in human's data, but a weak correlation for irregular data. In addition, with a closer examination of the verb by verb production probability, it seems that humans are more flexible in generating regular or irregular verbs than the models. In human's data, although the regular form appears to be dominant for most of the verbs, the various irregulars can still be produced even with such strong regular preference. The models lack such flexibility and produce the outputs in a more absolute manner. For example, the models output only the regular forms of many verbs and do not output any irregular forms. Similarly, there are also verbs that the models produce the vast majority of irregular forms. The models are more strongly influenced by their regular or irregular bias on each verb than humans.

6 Discussion and Conclusion

In this work, we demonstrate that the transformer models exhibit some abstract representation of regular and irregular verbs in past tense inflection generation. This abstract representation is largely affected by the type frequency of the input data. Since the regulars have a higher type frequency, the abstract representation is more robust for regular verbs than the irregular verbs. In addition, as long as the model could correctly classify the regular verb, it rarely makes errors in applying the correct inflection. Given the low type frequency and highly diverse inflection patterns for the irregular verbs, it is challenging for the model not only to classify the irregulars correctly, but also to apply the appropriate inflections. We found that increasing the type ratio would improve classification, and increasing token frequency would improve applying the correct inflections.

In addition, we also compared the model's nonce verb output with human data. The correlation with human data varies greatly for different models, which makes it difficult to state whether the neural models can capture human behavior. In our best-performing model, we observe that the model is able to produce both regular and irregular forms for a nonce verb. However, the models are more influenced by their own regular or irregular bias than human speakers. For example, humans can generate various forms even with a strong preference for regulars. However, the models are likely to generate either regular or irregular forms for a certain verb. Thus we conclude that the model's performance does not fit human data well.

Neural models have long been viewed as an approach against abstract representations. Therefore, the neural models are often rejected as cognitive models. In our work, we showed that the models exhibit some abstract representations, although still have a weak correlation with human performance for different reasons. We hope our findings could imply that the dual-route mechanism is not necessarily against each other and lead to more discussions about incorporating both sides of the debate to build a better cognitive model.

References

Adam Albright and Bruce Hayes. 2003. Rules vs. analogy in english past tenses: A computational/experimental study. *Cognition*, 90(2):119–161.

R Harald Baayen, Richard Piepenbrock, and Leon Gulikers. 1995. The celex lexical database (release 2). *Distributed by the Linguistic Data Consortium, University of Pennsylvania*.

Deniz Beser. 2021. Falling through the gaps: Neural architectures as models of morphological rule learning. *arXiv preprint arXiv:2105.03710*.

Lois Bloom. 1973. *One word at a time: The use of single word utterances before syntax*, volume 154. Walter de Gruyter.

Lois Bloom, Lois Hood, and Patsy Lightbown. 1974. Imitation in language development: If, when, and why. *Cognitive psychology*, 6(3):380–420.

Joan Bybee. 2003. *Phonology and language use*, volume 94. Cambridge University Press.

Maria Corkery, Yevgen Matusevych, and Sharon Goldwater. 2019. Are we there yet? encoder-decoder neural networks as cognitive models of english past tense inflection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3868–3877.

Christine Cuskley, Francesca Colaiori, Claudio Castellano, Vittorio Loreto, Martina Pugliese, and Francesca Tria. 2015. The adoption of linguistic rules in native and non-native speakers: Evidence from a wug task. *Journal of Memory and Language*, 84:205–223.

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.

Roy Patrick Higginson. 1985. *Fixing: Assimilation in language acquisition*. Ph.D. thesis, Washington State University.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Christo Kirov and Ryan Cotterell. 2018. Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.

Elena Lieven, Dorothé Salomo, and Michael Tomasello. 2009. Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, 20(3):481–507.

Brian MacWhinney. 2000. *The CHILDES project: The database*, volume 2. Psychology Press.

Gary F Marcus, Ursula Brinkmann, Harald Clahsen, Richard Wiese, and Steven Pinker. 1995. German inflection: The exception that proves the rule. *Cognitive psychology*, 29(3):189–256.

Gary F Marcus, Steven Pinker, Michael Ullman, Michelle Hollander, T John Rosen, Fei Xu, and Harald Clahsen. 1992. Overregularization in language acquisition. *Monographs of the society for research in child development*, pages i–178.

Kate McCurdy, Sharon Goldwater, and Adam Lopez. 2020. Inflecting when there's no majority: limitations of encoder-decoder neural networks as cognitive models for german plurals. *arXiv preprint arXiv:2005.08826*.

Steven Pinker and Alan Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(1-2):73–193.

David. E. Rumelhart and James L. McClelland. 1986. *On Learning the Past Tenses of English Verbs*, page 216–271. Cambridge, MA, USA.

Jacqueline Sachs. 1983. Talking about the there and then: The emergence of displaced reference in parent-child discourse. *Children's language*, 4:1–28.

Assaf Singer and Katharina Kann. 2020. The nyu-cuboulder systems for sigmorphon 2020 task 0 and task 2. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 90–98.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. *Advances in neural information processing systems*, 28.

Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. 2021. Frequency effects on syntactic rule learning in transformers. *arXiv preprint arXiv:2109.07020*.

Hao Zhang, Richard Sproat, Axel H Ng, Felix Stahlberg, Xiaochang Peng, Kyle Gorman, and Brian Roark. 2019. Neural models of text normalization for speech applications. *Computational Linguistics*, 45(2):293–337.

Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. 2019. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. In *Proceedings of NAACL-HLT*, pages 156–165.

7 Appendix

7.1 Data

7.1.1 Parents' Data

We created a dataset with parents' input past verbs with a higher irregular rate. We selected 8 children's corpora in the CHILDES database (MacWhinney, 2000) and aggregated their parents' past tense verbs. These 8 children include Adam, Eve, Sarah, Peter (Bloom, 1973), Allison (Bloom et al., 1974), Naomi (Sachs, 1983), April (Higginson, 1985), and Fraser (Lieven et al., 2009). All 8 children have been extensively studied in the previous literature to show that they have overregularization errors at an early age. However, we didn't use it as one of our training sets, because this dataset is too small for training from scratch, including only 411 unique past tense verbs with 69 unique irregulars (irregular verb ratio is 16.8%). If we leverage the percentage of its irregulars with the same construction method of TOKEN$_{irr}$, the dataset size would be 13,854 with an irregular ratio of 72.6%, which we used for the irregular ratio for REG$_{ds}$.

7.2 Data Cleaning

We cleaned the dataset used in KC (Kirov and Cotterell, 2018) by checking each verb's past tense in Merriam Webster dictionary and annotating the pronunciation of each verb with IPA. In KC's dataset, 14 verbs' past tenses and their labels are inconsistent, which are labeled with * in Table 14, and 2 verbs' past tenses are inconsistent with Merriam Webster dictionary, which are labeled with †. There are 33 verbs that have both regular and irregular past tense.

7.3 Accuracy by Verb Class

We report the test accuracy by verb class on regulars/irregulars of different models in Table 15 and Table 16.

7.4 Correlation

The correlations with human data for different models are listed in Table 17.

Verb	KC’s past tense	KC’s label	Merriam Webster
Verbs with both regular and irregular past tense			
abide	abided	reg	abided, abode
alight	alighted	reg	alighted, alit
awake	awoke	irreg	awoke, awaked
beseech	besought	irreg	beseeched, besought
bet	betted	irreg*	bet, betted
broadcast	broadcasted	reg	broadcast, broadcasted
cleave	cleaved	reg	cleaved, clove, clave
clothe	clothed	reg	clothed, clad
dive	dived	irreg*	dived, dove
dream	dreamed	irreg*	dreamed, dreamt
floodlight	floodlighted	reg	floodlit, floodlighted
gild	gilded	reg	giled, gilt
gird	girded	reg	girded, girt
hang	hung	irreg	hung, hanged
inset	insetted	irreg*	inset, insetted
knit	knitted	irreg*	knit, knitted
leap	leaped	irreg*	leaped, leapt
light	lighted	irreg*	lit, lighted
outshine	outshone	irreg	outshone, outshined
plead	pleaded	reg	pleaded, pled
quit	quitted	irreg*	quit, quitted
rend	rent	reg*	rent, rended
shine	shone	irreg	shone, shined
shoe	shod	reg*	shod, shoed
sneak	sneaked	irreg*	sneaked, snuck
speed	speeded	irreg*	sped, speeded
spit	spat	irreg	spit, spat, spitted
stick	stuck	irreg	sticked, stuck
strive	strove	irreg	strove, strived
sweat	sweated	reg	sweat, sweated
tread	trod	irreg	trod, treaded
wed	wedded	reg	wedded, wed
wet	wetted	irreg*	wet, wetted
Verbs with more than one irregular past tense.			
beget	begot	irreg	begot, begat
bid	bade	irreg	bade, bid
sing	sang	irreg	sing, sung
sink	sank	irreg	sank, sunk
KC’s data inconsisted with Merriam Webster			
cost	costed†	irreg*	cost
shit	shitted†	reg	shit, shat

Table 14: The verbs and their past tense listed in KC’s dataset and Merriam Webster dictionary. *indicates that the KC’s label and its past tense do not match. † indicates the past tense in KC is not listed in the dictionary.

Train Set	Model	/-d/		/-t/		/ɪd/	
		van.	copy	van.	copy	van.	copy
TYPE$_{reg}$	BASE	100	100	94	98	94	97
	LABEL$_{reg}$	100	99	98	96	98	97
	LABEL$_{vc}$	99	100	97	97	97	97
	LABEL$_2$	98	99	97	99	98	96
TYPE$_{irr}$	BASE	99	100	96	98	92	98
	LABEL$_{reg}$	100	98	100	98	94	96
	LABEL$_{vc}$	98	99	95	94	99	98
	LABEL$_2$	99	99	96	94	97	93
TOKEN$_{both}$	BASE	95	99	97	99	100	98
	LABEL$_{reg}$	96	96	98	97	99	100
	LABEL$_{vc}$	98	94	95	98	98	99
	LABEL$_2$	98	95	99	98	97	100
TOKEN$_{irr}$	BASE	95	95	93	99	98	98
	LABEL$_{reg}$	94	92	98	97	95	96
	LABEL$_{vc}$	94	90	96	95	97	96
	LABEL$_2$	92	93	94	96	96	95

Table 15: Test accuracy (%) of different models on regulars by verb class.

		vc		vc+/-t/		vc+/-d/		ruck		weak		level		other	
		van.	copy	van.	copy	van.	copy	van.	copy	van.	copy	van.	copy	van.	copy
TYPE$_{reg}$	BASE	6.7	0	6.7	0	0	6.7	0	0	6.7	0	0	0	13.3	6.7
	LABEL$_{reg}$	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	LABEL$_{vc}$	0	0	0	0	0	0	0	0	0	0	0	0	13.3	13.3
	LABEL$_2$	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TYPE$_{irr}$	BASE	6.7	6.7	6.7	0	0	0	0	0	6.7	6.7	0	6.7	6.7	0
	LABEL$_{reg}$	0	0	0	0	0	0	0	0	0	0	0	6.7	0	0
	LABEL$_{vc}$	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	LABEL$_2$	0	0	0	0	6.7	0	0	0	0	0	0	0	6.7	0
TOKEN$_{both}$	BASE	0	0	13.3	13.3	26.7	20.0	26.7	6.7	0	0	6.7	13.3	20.0	33.3
	LABEL$_{reg}$	0	0	20.0	0	13.3	13.3	6.7	0	13.3	0	0	6.7	0	6.7
	LABEL$_{vc}$	0	0	13.3	0	13.3	0	0	0	0	0	13.3	0	6.7	20.0
	LABEL$_2$	0	0	20.0	13.3	0	0	0	0	0	0	0	0	0	6.7
TOKEN$_{irr}$	BASE	13.3	13.3	40.0	13.3	40.0	6.7	26.7	0	13.3	6.7	26.7	6.7	33.3	33.3
	LABEL$_{reg}$	0	0	20.0	20.0	40.0	13.3	6.7	0	13.3	20.0	13.3	6.7	6.7	0
	LABEL$_{vc}$	6.7	0	20.0	20.0	0	6.7	0	0	13.3	0	0	0	13.3	6.7
	LABEL$_2$	0	6.7	6.7	20.0	20.0	6.7	0	0	13.3	6.7	6.7	0	0	6.7

Table 16: Test accuracy (%) of different models on irregulars by verb class.

No Copy Mechanism		Regular (N = 58) $P_{pro} r$	Rate ρ	Irregular (N = 58) $P_{pro} r$	Rate ρ	Irregular 2 (N = 16) $P_{pro} r$	Rate ρ
TYPE$_{reg}$	BASE	0.01	0.28	0.62	0.47	0.01	0.34
	LABEL$_{reg}$	-0.14	0.23	0.33	0.28	0.20	-0.02
	LABEL$_{vc}$	-0.13	0.05	0.06	-0.06	0.28	0.42
	LABEL$_2$	-0.04	0.28	0.43	0.17	0.23	0.14
TYPE$_{irr}$	BASE	-0.15	0.02	0.31	0.34	NaN	NaN
	LABEL$_{reg}$	-0.02	0.46	0.36	0.28	-0.23	0.01
	LABEL$_{vc}$	-0.05	0.28	0.26	0.21	0.33	0.05
	LABEL$_2$	-0.02	0.48	0.40	0.37	0.56	0.13
TOKEN$_{both}$	BASE	0.57	0.51	0.29	0.33	0.33	0.11
	LABEL$_{reg}$	0.48	0.43	0.26	0.30	-0.25	0.12
	LABEL$_{vc}$	0.57	0.59	0.22	0.22	0.12	0.36
	LABEL$_2$	0.42	0.41	0.19	0.17	-0.13	0.09
TOKEN$_{irr}$	BASE	0.26	0.36	0.19	0.13	0.39	0.13
	LABEL$_{reg}$	0.24	0.41	0.23	0.22	-0.16	-0.02
	LABEL$_{vc}$	0.27	0.40	0.18	0.21	-0.17	0.14
	LABEL$_2$	0.30	0.43	0.20	0.14	-0.05	-0.04
Copy Mechanism							
TYPE$_{reg}$	BASE	-0.14	0.20	0.12	0.30	NaN	0.45
	LABEL$_{reg}$	-0.07	0.31	0.22	0.41	NaN	NaN
	LABEL$_{vc}$	-0.11	0.28	NaN	0.44	NaN	-0.03
	LABEL$_2$	-0.16	0.32	NaN	0.36	NaN	0.10
TYPE$_{irr}$	BASE	-0.04	0.36	0.29	0.51	0.64	0.08
	LABEL$_{reg}$	-0.19	0.29	0.30	0.51	NaN	-0.25
	LABEL$_{vc}$	-0.12	0.33	0.23	0.55	NaN	0.14
	LABEL$_2$	-0.17	0.15	NaN	0.40	NaN	NaN
TOKEN$_{both}$	BASE	0.36	0.28	0.17	0.18	0.33	0.07
	LABEL$_{reg}$	0.35	0.32	0.23	0.21	0.26	0.08
	LABEL$_{vc}$	0.14	0.30	0.12	0.26	-0.25	-0.06
	LABEL$_2$	0.18	0.16	0.13	0.08	-0.04	0.05
TOKEN$_{irr}$	BASE	0.30	0.27	0.26	0.30	0.29	0.09
	LABEL$_{reg}$	0.23	0.24	0.13	0.21	-0.12	0.08
	LABEL$_{vc}$	0.27	0.32	0.16	0.22	-0.25	0.04
	LABEL$_2$	0.34	0.41	0.14	0.31	0.18	0.04
Resembing Methods Without Copy Mechanism							
BALANCE	BASE	0.62	0.74	0.44	0.45	0.69	0.28
	LABEL$_{reg}$	0.57	0.74	0.44	0.35	0.06	0.22
	LABEL$_{vc}$	0.63	0.70	0.47	0.42	0.42	0.33
	LABEL$_2$	0.64	0.79	0.43	0.31	0.35	0.24
REG$_{ds}$	BASE	0.61	0.74	0.46	0.51	0.55	0.11
	LABEL$_{reg}$	0.61	0.66	0.51	0.39	0.08	0.24
	LABEL$_{vc}$	0.48	0.60	0.42	0.40	0.49	0.51
	LABEL$_2$	0.50	0.65	0.40	0.31	0.15	0.41
IRREG$_{ds}$	BASE	0.68	0.74	0.52	0.52	0.08	0.06
	LABEL$_{reg}$	0.52	0.63	0.39	0.33	0.77	0.43
	LABEL$_{vc}$	0.70	0.77	0.44	0.28	0.58	0.31
	LABEL$_2$	0.49	0.63	0.39	0.34	0.39	0.09
Resembing Methods With Copy Mechanism							
BALANCE	BASE	0.54	0.70	0.34	0.32	0.48	0.11
	LABEL$_{reg}$	0.51	0.69	0.49	0.39	0.42	0.40
	LABEL$_{vc}$	0.65	0.75	0.31	0.23	0.30	0.36
	LABEL$_2$	0.52	0.63	0.45	0.42	0.50	0.30
REG$_{ds}$	BASE	0.52	0.66	0.38	0.34	0.53	0.26
	LABEL$_{reg}$	0.54	0.67	0.37	0.35	0.28	0.34
	LABEL$_{vc}$	0.50	0.69	0.49	0.43	0.43	0.31
	LABEL$_2$	0.51	0.67	0.31	0.21	0.40	0.15
IRREG$_{ds}$	BASE	0.60	0.73	0.46	0.39	0.38	0.35
	LABEL$_{reg}$	0.63	0.76	0.42	0.34	0.56	0.48
	LABEL$_{vc}$	0.64	0.71	0.34	0.20	0.46	0.18
	LABEL$_2$	0.59	0.67	0.38	0.31	0.62	0.16

Table 17: Correlation with human data for different models. NaN represents the correlation that can not be computed due to too many zeros.