Evaluating Dialect Robustness of Language Models via Conversation Understanding

Dipankar Srirag[1], **Nihar Ranjan Sahoo**[2], **Aditya Joshi**[1]

[1]University of New South Wales, Sydney, Australia
[2]Indian Institute of Technology Bombay, India
{d.srirag, aditya.joshi}@unsw.edu.au, nihar@cse.iitb.ac.in

Abstract

With an evergrowing number of LLMs reporting superlative performance for English, their ability to perform equitably for different dialects of English (*i.e.*, dialect robustness) needs to be ascertained. Specifically, we use English language (US English or Indian English) conversations between humans who play the word-guessing game of 'taboo'. We formulate two evaluative tasks: target word prediction (TWP) (*i.e.*, predict the masked target word in a conversation) and target word selection (TWS) (*i.e.*, select the most likely masked target word in a conversation, from among a set of candidate words). Extending MD3, an existing dialectic dataset of taboo-playing conversations, we introduce M-MD3, a target-word-masked version of MD3 with the en-US and en-IN subsets. We create two subsets: en-MV (where en-US is transformed to include dialectal information) and en-TR (where dialectal information is removed from en-IN). We evaluate one open-source (Llama3) and two closed-source (GPT-4/3.5) LLMs. LLMs perform significantly better for US English than Indian English for both TWP and TWS tasks, for all settings, exhibiting marginalisation against the Indian dialect of English. While GPT-based models perform the best, the comparatively smaller models work more equitably after fine-tuning. Our error analysis shows that the LLMs can understand the dialect better after fine-tuning using dialectal data. Our evaluation methodology exhibits a novel way to examine attributes of language models using pre-existing dialogue datasets.

1 Introduction

Large language models (LLMs)[1] based on Transformers (Vaswani et al. 2017) are the state-of-the-art in natural language processing (NLP), often reporting superlative performance on several NLP tasks (Zhao et al. 2023). These models predominantly use English language data in their pre-training corpus. However, being a widely spoken language, English takes multiple forms in different parts of the world. These forms, called dialects or national varieties of English, collectively constitute the World Englishes (Bolton 2012). While research papers introducing LLMs report performance on English language datasets, recent works highlight the performance gap between US English and other dialects of English for several natural language processing tasks (Joshi et al. 2024).

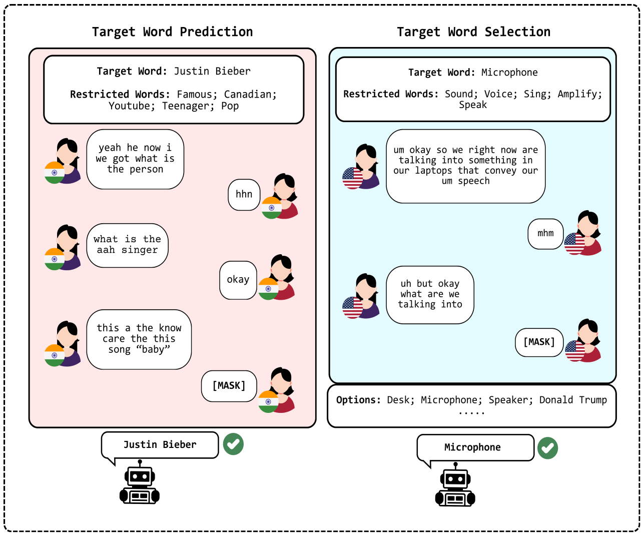
Target Word Prediction
Target Word Selection
Target Word: Justin Bieber
Restricted Words: Famous; Canadian; Youtube; Teenager; Pop
yeah he now i we got what is the person
hhn
what is the aah singer
okay
this a the know care the this song "baby"
[MASK]
Justin Bieber
Target Word: Microphone
Restricted Words: Sound; Voice; Sing; Amplify; Speak
um okay so we right now are talking into something in our laptops that convey our um speech
mhm
uh but okay what are we talking into
[MASK]
Options: Desk; Microphone; Speaker; Donald Trump .....
Microphone

Figure 1: Illustration of the two tasks: Target word prediction (TWP) and Target word selection (TWS). 👤and 👤are the describer and the guesser respectively in a word-guessing game of taboo. 🇮🇳and 🇺🇸refer to Indian English and US English respectively.

Our paper follows this line of work of evaluating LLMs for dialects of English via conversation understanding. The choice of conversation understanding as a domain for evaluation emerges from the fact that dialectal features are most visible in free-flowing conversations (Negro and Vietti 2006). Therefore, we investigate the research question:

> "*In comparison with US English, how effectively can LLMs understand conversations between speakers of other national varieties of English?*"

To address the research question, we use the **MD3** (Eisenstein et al. 2023) that consists of manually transcribed dialogues between pairs of human participants where each pair speaks either Indian English or US English. The participants engage in a focused conversation: they play the word-guessing game based on the game of 'Taboo' (Wikipedia

[1] We use 'language models' and 'large language models/LLMs' interchangeably in this paper.

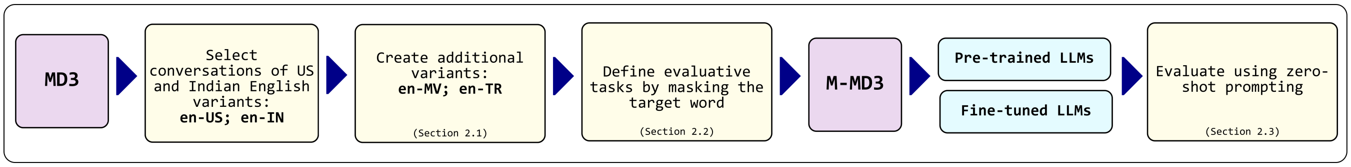
MD3
Select conversations of US and Indian English variants: en-US; en-IN
Create additional variants: en-MV; en-TR
(Section 2.1)
Define evaluative tasks by masking the target word
(Section 2.2)
M-MD3
Pre-trained LLMs
Fine-tuned LLMs
Evaluate using zero-shot prompting
(Section 2.3)

Figure 2: Steps for evaluation of dialect robustness.

2023). In the game, a describer must get a guesser to identify a **target word** but must not use a set of related words known as **restricted words** while describing the target word. Using this dataset of dialectal dialogues, we introduce two tasks to evaluate the dialect-robustness of LLMs to understand conversations. They are: (a) Given an input conversation with the target word masked, can the LLM *predict* the target word? (referred to as **target word prediction**) (b) Given an input conversation with the target word masked along with a set of candidate target words, can the LLM *select* the correct target word? (referred to as **target word selection**). Our approach of masking the target word is similar to Dey and Desarkar (2023), who show that masked word prediction may correlate with automatic dialogue evaluation metrics. Figure 1 shows an example of the two tasks, where the language model predicts '*Justin Bieber*' for target word prediction, and selects '*microphone*' among the set of options for target word selection[2]. For the two tasks, we extend MD3 to create a target-word-**M**asked **M**ulti-**D**ialect **D**ataset of **D**ialogues (**M-MD3**)[3]. M-MD3 consists of (a) conversations between Indian English speakers (en-IN), and conversations between US English speakers (en-US), (b) en-US conversations transformed into en-IN using rule-based perturbations (en-MV), (c) en-IN with dialectal information removed (en-TR). We evaluate the performance of three SOTA large language models (LLMs), one open-source and two closed-source, employing zero-shot prompting on both pre-trained and fine-tuned models (where available).

Our evaluation methodology derives from past work that evaluates LLMs by providing a set of task-specific examples (Wang et al. 2023). Of particular relevance is the work by Chalamalasetti et al. (2023), who generate word game conversations using LLMs and evaluate their ability to predict the target word. The contributions of our work are:

- We create M-MD3, an extension of MD3, that deals with two novel evaluative tasks for dialect robustness: target word prediction and target word selection.
- Our evaluation demonstrates a degraded performance in the case of Indian English as compared to US English for all models, supporting existing social disparities between US and Indian culture in the LLM representations (Khandelwal et al. 2024).
- A comprehensive error analysis to identify specific conditions under which fine-tuning enhances the model's performance on Indian English conversations.

Since several LLMs have been deployed as publicly available dialogue agents[4], it is imperative that they are able to understand conversations for users belonging to diverse English-speaking subgroups. In the case of our paper, this refers to dialectal variations. The rest of the paper is organized as follows. Section 2 introduces our evaluation methodology. The experiment setup and results are in Sections 3 and 4 respectively.

2 Methodology

We present our method step-by-step, with a detailed overview of our evaluation methodology described in Figure 2. We select two subsets available in MD3: en-IN and en-US, and filter out the conversations where the guesser could not identify the target word. We extend MD3 to include two additional sets of conversations—en-MV and en-TR, and mask the target words in all four subsets to create M-MD3. We ensure that the mask token always appears at the end of the conversation, warranting the use of auto-regressive models. This is done by pruning the conversation to the turn where the guesser utters the target word[5].

2.1 Extending MD3

Transforming text in en-US to other dialectal English text has been explored for low-resource settings (Held, Ziems, and Yang 2023; Xiao et al. 2023; Liu, Held, and Yang 2023). To evaluate the efficacy of synthetically transformed dialogues, we extend the dataset of dialectal dialogues to include two additional sets of conversations– en-MV and en-TR.

en-MV We use Multi-VALUE (Ziems et al. 2023) to transform en-US conversations into en-IN conversations. We call this set of conversations created by rule-based transformations en-MV.

en-TR We prompt GPT-4 Turbo Preview(GPT-4; OpenAI 2024) to remove dialectal information from en-IN. The resultant set of conversations is known as en-TR. The prompt used to generate such conversations is given below:

> "*Normalise the conversation. Remove all exaggerations and dialectal information. Return a neutral response.*"

[2] We run experiments on both the tasks for both US and Indian English conversations. While the examples show expected output, the LLM may or may not produce the same in the case of our experiments. That is the crux of the evaluation.

[3] M-MD3 dataset and the related code will be made publicly available at https://github.com/dipankarsrirag/eval-dialect-robust.

[4] ChatGPT https://chat.openai.com/; Accessed on 9th April 2024.

[5] Details on the masking method with examples are provided in the *Dataset Construction* section of the Supplementary material.

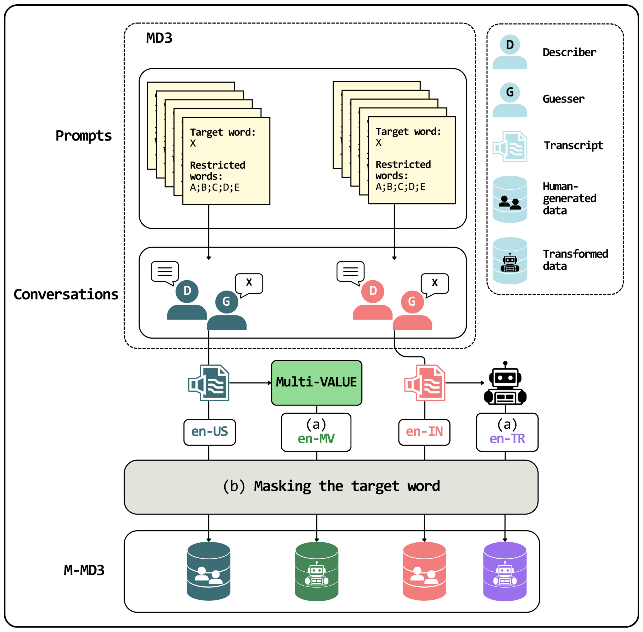

Figure 3: M-MD3 as an extension of MD3: (a) Creation of en-MV and en-TR, and (b) Creation of target-word-masked conversations.

The use of GPT-4 to transform en-IN conversations sometimes leads to the generation of conversation summaries rather than transformed conversations[6]. Due to the varying lengths of speaker turns, transforming en-US conversations using Multi-VALUE occasionally fails to output a result. Such failed transformations are excluded from both the subsets of transformed (en-MV, en-TR) conversations, leading to fewer conversations in en-MV and en-TR as compared to en-US and en-IN, respectively, as shown in Tables 1 and 2.

Subset	Avg. turns	Avg. words	Single	Multiple
en-US	4.1	42.1	308	106
en-IN	6.8	57.4	153	59
en-MV	4.9	35.2	245	87
en-TR	6.3	42.7	121	50

Table 1: Constructional Statistics of M-MD3. *Single* and *Multiple* refer to the number of conversations with single-word and multiple-word reference targets, respectively.

2.2 Analysis

Table 1 reports some of the constructional statistics of M-MD3. For each subset, it reports the average number of dialogue turns per conversation, the average word count for the dialogues uttered by both the describer and the guesser, and the number of conversations with single-word versus multiple-word reference target words. The target words '*microphone*' and '*Justin Bieber*' in Figure 1 are examples of single-word and multiple-word reference target words, respectively.

We notice a higher number of average turns and words spoken in en-IN conversations compared to en-US conversations. This is due to the en-US speakers being more familiar with the target word compared to en-IN speakers, leading to shorter gameplay time (Eisenstein et al. 2023). The trend is also carried over to the transformed conversations in en-MV (derived from en-US) and en-TR (derived from en-IN).

2.3 Task Definition

As shown in Figure 3, we mask the target word in the conversations from all four subsets. The target word occurs in the last dialogue turn of the conversation, which is spoken by the guesser[7]. As a result, we formulate two tasks where the expected output is to fill the correct word at the masked position:

- **Target Word Prediction (TWP)**: Given a conversation with the target word masked, predict the word.
- **Target Word Selection (TWS)**: Given a conversation with the target word masked and a set of candidate target words, select one among the candidate set.

In the case of TWP, the LLM may generate any word within its learned vocabulary, with the expected output being the reference target word. In the case of TWS, we provide the LLM with a masked conversation and a set of all target words in the M-MD3 corpus. The LLM must then select the most likely target word.

We then use prompting on three LLMs to perform both tasks (TWP and TWS). As LLMs, we choose models that have been optimised to follow natural language instructions. In our case, the instruction is to either predict the masked target word or select a word from candidate words. Specifically, we use one open-source model, namely, Llama 3 70B Chat (LLAMA-3; Llama Team 2024), and two closed-source models, namely, GPT-4 and GPT-3.5 Turbo 0125 (GPT-3.5; Ouyang et al. 2022).

3 Experiment Setup

We report the performance on pre-trained and fine-tuned versions of LLMs using zero-shot prompting. Fine-tuning is always done '*in-dialect*' in our case, although there is no reason to believe that cross-dialect fine-tuning is not possible.

3.1 Model Parameters

Experiments on GPT-4 and GPT-3.5 are conducted using OpenAI's API[8]. GPT-3.5 is fine-tuned for 5 epochs, separately for every subset. We select top_p as 0.2 to restrict variability in output generation.

[6] More details with examples are discussed in the *Transformation Issues* section of the Supplementary material.

[7] This always holds because of the way we process the conversations.

[8] OpenAI API https://platform.openai.com/docs/api-reference; Accessed on 18th April 2024.

LLAMA-3 is fine-tuned for 20 epochs, with a batch size of 16, Paged 8-bit AdamW (Dettmers et al. 2022) as the optimiser and a learning rate of 2e-4. We use QLoRA adaptors, targeting all linear layers, as recommended by Dettmers et al. (2023). All experiments on LLAMA-3 were performed using two A100 GPUs.

3.2 Metrics

We report our results on two metrics: *accuracy* and *similarity*. *Accuracy* is the proportion of conversations where the LLM generated the correct target word. This is a strict metric in that it requires the LLM to generate an exact match to the reference target word. In the case of TWP, the LLM will choose from all the words within its vocabulary, while in the case of TWS, the LLM will choose from the set of candidate target words. Therefore, it is trivial that the accuracy for TWS is expected to be higher than that for TWP. Accuracy metric penalizes models even if the generated target word partially matches with the reference target word in case of multi-word reference target as described in Section 2.2. As *similarity*, we report the cosine similarity between the Sentence-BERT embeddings (Reimers and Gurevych 2019) of the reference target word and the generated target word. This allows for similar but inexact words generated by the LLM to be acceptable to the similarity score.

3.3 Experiments

We perform experiments on both the tasks (TWP and TWS) using all models($\{$pre-trained and fine-tuned$\}$ × $\{$GPT-4, GPT-3.5, LLAMA-3 $\}$). All results are reported only on the test split of each subset of conversations. All fine-tuned models are fine-tuned on the training and validation set using instruction fine-tuning. GPT-4 could not be fine-tuned because doing so is restricted by OpenAI at the time of writing this paper. The statistics of **Train**, **Valid**, and **Test** splits of each subset of M-MD3 are reported in Table 2.

Subset	Train	Valid	Test
en-US	62	41	311
en-IN	31	21	160
en-MV	49	33	250
en-TR	23	17	131

Table 2: Statistics of M-MD3.

4 Results

In this section, we compare the performance of three LLMs both quantitatively and qualitatively. Note that the same test split is used to evaluate both pre-trained and fine-tuned versions, ensuring that the results are comparable.

4.1 Quantitative Results

Table 3 shows the results of our experiments on each task specified in Section 2.3. We analyse the results as follows.

en-US versus en-IN The focus of this paper is to evaluate dialect robustness by comparing the performance on en-US and en-IN. All LLMs perform consistently better on en-US as compared to en-IN for all configurations. For example, in the case of LLAMA-3 and TWP, the similarity scores on the fine-tuned model are 78.0 for en-US and 66.3 for en-IN, with the drop in performance of 11.7. Even for all three models, en-US outperforms en-IN on zero-shot performance using the pre-trained model. From all results, it is clearly understood that, on average, the LLMs understand the US English dialect better than the Indian English dialect. Only considering the pre-training setting, GPT-4 outperforms other models for both en-US and en-IN. However, fine-tuning improves the performance of LLAMA-3 on en-IN, achieving better results on both tasks compared to GPT-based models. Interestingly, for LLAMA-3, the performance improvement after fine-tuning on en-IN is greater compared to fine-tuning on en-US (represented by Δ).

Impact of transforming conversations As discussed in section 2.1, we introduced two synthetically transformed subsets, en-MV and en-TR, to assess the importance of dialectal features in LLMs' understanding of conversations. Table 3 shows that, on pre-trained models, en-TR conversations have better performance compared to original en-IN conversations. This suggests that after removing the dialectal information from en-IN, the resulting en-TR conversations are close to the distribution of the dialect that the LLM understands. This behaviour is better reflected in GPT-3.5, potentially, because the LLM has a poor understanding of en-IN as compared to the other two LLMs. Additionally, fine-tuning on en-TR conversations does not improve the task performances in comparison to that on en-IN. This supports the hypothesis that the removal of dialectal information brings the resulting conversation closer to the dialectal distribution that LLMs understand than the original dialect.

In the case of en-MV, the task performances are consistently lower compared to en-US. For example, in the case of GPT-3.5 and TWS, the similarity scores on the fine-tuned model are 80.8 for en-US and 71.5 for en-MV. This degraded performance shows that the rule-based transformation into en-IN from en-US reduced the understanding capacity of LLMs for the resulting conversations, further strengthening our hypothesis that LLMs perform well for US English dialects compared to any other varieties, similar to findings of Ryan, Held, and Yang (2024).

Shorter turns versus Longer turns A trend appears between the performances of models on each subset of conversations and the constructional properties of these conversations discussed in Section 2.2. Models report their best performances on the subset with the smallest number of average turns in a conversation (en-US), and report the worst performance on the subset with the highest number of average turns in a conversation (en-IN).

TWP versus TWS We now compare the performances of TWP and TWS. As expected, the similarity and accuracy are higher in the case of TWS compared to TWP for all three models, with one exception: the pre-training performance of

Model	Subset	TWP						TWS					
		Similarity			Accuracy			Similarity			Accuracy		
		PT	FT	Δ	PT	FT	Δ	PT	FT	Δ	PT	FT	Δ
GPT-4	en-US	77.4	–	–	67.8	–	–	85.7	–	–	78.8	–	–
	en-IN	63.0	–	–	45.6	–	–	79.0	–	–	72.5	–	–
	en-MV	75.6	–	–	60.0	–	–	83.6	–	–	74.4	–	–
	en-TR	62.8	–	–	45.8	–	–	83.4	–	–	77.1	–	–
	δ	-14.4	–	–	-22.0	–	–	-6.7	–	–	-6.3	–	–
GPT-3.5	en-US	66.3	72.2	5.9	52.7	59.1	6.4	66.4	80.8	14.4	50.8	71.3	20.5
	en-IN	53.2	59.1	5.9	34.4	40.0	5.6	61.9	70.7	8.8	47.5	60.6	13.1
	en-MV	57.6	71.3	13.7	40.0	54.4	14.4	52.4	71.5	19.1	31.6	57.6	26.0
	en-TR	59.4	61.0	1.6	39.7	41.2	1.5	70.7	73.0	2.3	57.3	60.3	3.0
	δ	-13.1	-13.1	–	-18.3	-19.1	–	-4.5	-10.1	–	-21.0	-16.2	–
LLAMA-3	en-US	70.8	78.0	7.2	60.5	65.3	4.8	78.0	81.8	3.8	67.5	74.6	7.1
	en-IN	59.8	66.3	6.5	43.8	54.4	10.6	68.8	80.8	12.0	56.9	74.4	17.5
	en-MV	68.6	73.8	5.2	54.0	61.6	7.6	72.3	77.6	5.3	58.8	67.2	8.4
	en-TR	60.7	57.5	-3.2	45.8	42.7	-3.1	70.8	79.5	8.7	60.3	72.5	12.2
	δ	-11.0	-11.7	–	-16.7	-10.9	–	-9.2	-1.8	–	-10.6	-0.2	–

Table 3: Performance on the two tasks: TWP and TWS. PT/FT: Pre-trained/Fine-tuned. δ is the difference in performance between en-IN and en-US (en-IN minus en-US). Δ is the difference in performance between FT and PT.

GPT-3.5 on en-MV, where TWP slightly outperforms TWS. Note that, for pre-trained LLAMA-3, the accuracy on en-IN is 43.8 for TWP and 56.9 for TWS. Across all configurations, fine-tuning consistently improves the performance of both TWP and TWS. GPT-4 performs best (only for pre-trained models) for both TWP and TWS tasks for all subsets.

Model Comparison It can be easily observed from Table 3 that the GPT-4 outperforms the other two LLMs in the pre-training setting. Interestingly, for TWS, GPT-4 pre-training performances are better than fine-tuning performances of GPT-3.5 and LLAMA-3 in most of the cases. Also, GPT-4 performs almost equally well for each subset of M-MD3. This shows that GPT-4 and LLAMA-3 are more inclusive for different dialectal variations of English in the pre-training and fine-tuning setting, respectively.

Pre-training versus Fine-tuning Although the pre-training performances of GPT-4 are superlative, Table 3 shows that the fine-tuning also improves the performance of GPT-3.5 and LLAMA-3 across both tasks and four subsets. Fine-tuning is more effective for en-US than en-IN in the case of GPT-3.5, whereas LLAMA-3 shows the opposite trend. For GPT-3.5, the most improvement due to fine-tuning is seen when the models are fine-tuned on en-MV, while LLAMA-3 shows the highest improvement when fine-tuned on en-IN.

4.2 Error Analysis

From **Test** set of each conversation subset, we randomly select 30 conversations that are mislabeled by GPT-4 and LLAMA-3, and manually analyse errors among all model variants across all subsets of conversations. We summarise the six error categories[9] in Table 4. The error types are:

Ambigous Descriptions (AD) This error type is observed when descriptions lack specificity (given the *situational* constraint on the describer), leading to multiple potential answers. For the example target word–'*engine*,' the description provided is–'*What we find in our. cars. in the front part?*'. Although these descriptions provide enough information to guide a human guesser to the right answer, they are often too vague to guide the LLM to a singular, correct interpretation.

Wrong Descriptions (WD) These errors occur when the guesser guesses the target word even before the describer can finish the description completely. In the case of the target word '*surname*,' the model infers '*parent*' when the description provided is–'*beside your. uh. what is your elder? Uh what is*'. While human guessers might use their cognitive bias to guess correctly without the complete description, LLMs lack the ability to understand the target word from such a description.

Broken down description of prompt word (BDD) This error occurs when the describer breaks down the target word into subwords and attempts to explain each separately. Generally, such descriptions involve longer turns. The guesser is then expected to piece together these fragments to deduce the original word, as in the case of the target word '*Billie Holiday*,' the describer individually describes the subwords '*Billie*' and '*Holiday*'. In such cases, LLMs sometimes latch

[9] Additonal examples for each error category are in the *Errors* section of the Supplementary material.

Error Type	Config	GPT-4				LLAMA-3			
		en-US	en-IN	en-MV	en-TR	en-US	en-IN	en-MV	en-TR
AD	PT	18 (5)	13 (3)	13 (6)	14 (6)	10 (6)	16 (10)	18 (15)	11 (7)
	FT	– (–)	– (–)	– (–)	– (–)	7 (6)	9 (4)	13 (13)	11 (6)
WD	PT	4 (2)	4 (4)	– (–)	3 (3)	3 (2)	5 (5)	– (–)	3 (3)
	FT	– (–)	– (–)	– (–)	– (–)	2 (2)	5 (4)	– (–)	3 (2)
BDD	PT	3 (2)	16 (5)	– (–)	7 (3)	– (–)	2 (1)	– (–)	3 (2)
	FT	– (–)	– (–)	– (–)	– (–)	– (–)	0 (0)	– (–)	2 (3)
CC	PT	6 (2)	5 (2)	12 (5)	4 (2)	4 (2)	5 (4)	5 (3)	2 (1)
	FT	– (–)	– (–)	– (–)	– (–)	2 (1)	4 (2)	3 (2)	2 (0)
PF	PT	6 (2)	2 (0)	7 (4)	4 (0)	14 (8)	3 (1)	9 (5)	4 (1)
	FT	– (–)	– (–)	– (–)	– (–)	7 (5)	1 (1)	5 (4)	3 (0)
ERR	PT	– (–)	– (–)	6 (1)	– (–)	– (–)	– (–)	4 (3)	– (–)
	FT	– (–)	– (–)	– (–)	– (–)	– (–)	– (–)	3 (1)	– (–)
$\sum$	PT	37 (13)	40 (14)	38 (16)	32 (14)	31 (18)	31 (21)	40 (29)	23 (14)
	FT	– (–)	– (–)	– (–)	– (–)	18 (13)	19 (11)	27 (21)	21 (11)

Table 4: Count of errors of GPT-4 and LLAMA-3 for each subset. PT/FT: Pre-trained/Fine-tuned. ‘X (Y)’ indicates that there are X errors in TWP and Y errors in TWS. $\sum$ is the sum of errors tagged in the sampled erroneous conversations by a model on a subset across all error types.

onto the descriptions pertaining to later subwords, predicting a partially correct target word.

Shared Cultural Context (CC) These errors arise when the human players use culturally shared notions in a conversation, often due to the describer's lack of familiarity with the target word. For example, an Indian describer explains the word ‘*idli*’ using examples of breakfast items and then asks the guesser to infer ‘*Adele*’. The model is unable to understand this happening in the conversation.

Public Figure (PF) These errors pertain to inaccurate predictions generated by the model when the descriptions are about a well-known public figure. For example, the describer describes the target word ’*Mike Tyson*’ as ‘*Big guy that punched people out and he had a little bit of a lisp*,’ but the model generates ‘*darth*’.

Fallback Error (ERR) While efforts were made to classify every mislabeled conversation into an error category, few generated target words were found to be inexact or inaccurate, even with apt descriptions in the conversations. For example, the target word–‘*Rose*’ and the description–‘*This are the types of that's often given valentine day plant.*’, the model generates ‘*Gift*’. This example description mentions the word *plant* which should have guided the model to a more specific target word than *Gift*.

The error types **AD**, **CC**, and **PF** test the model's ability to predict the target word based on descriptions influenced by the describer's dialect, shared notions with the guesser, and perceived notions about the target word. Also, some of the conversations fall into multiple error categories except in the case of conversations in **ERR** (which is a mutually exclusive label).

Table 4 presents the error cases in ‘X (Y)’ which indicates that there are X errors in TWP and Y errors in TWS for the corresponding configuration. The benefit of TWS providing options for the target word is seen in **AD**, where the alleviation is almost uniform across all dialects. The presence of direct or indirect references to the prompt word helps the LLM towards a plausible answer, in turn making it easier for them to choose an option. However, this error reduction does not extend to **CC**, which LLMs are unable to detect.

Fine-tuning helps to reduce the errors of **AD** category more for conversations of en-IN dialect compared to en-US. However, after removing the dialectal information, the conversations are insensitive to fine-tuning for the **AD** error cases. Additionally, fine-tuning helps to decrease errors in the **PF** category. As expected, it does not significantly reduce errors in the **WD** category.

5 Related Work

Research in **dialect robustness** stems from the need for language technologies to be equitable and not reinforce any negative sentiments against a specific linguistic subgroup (Blodgett et al. 2020). LLMs perform poorly on several downstream tasks (such as the tasks in the GLUE benchmark) involving dialects other than mainstream US English (Joshi et al. 2024; Faisal et al. 2024).

Similar to our work, the evaluation of language understanding ability of LLMs has been explored using typical **conversation understanding tasks** (Chen et al. 2022) like conversation summarisation (Gliwa et al. 2019; Chen et al. 2021), conversation completion (Sai et al. 2020; Ueyama and Kano 2023), or NLU tasks (Faisal et al. 2024). Other approaches involve conversation-based question-answering tasks that also evaluate the reasoning abilities of LLMs (Sun et al. 2019; Qin et al. 2021). Tasks like mask-filling were used to evaluate LLM-generated responses, more specifically Dey and Desarkar (2023) do so by making RoBERTa predict masked keyword utterances when given a context of dialogue history along with conditions like persona, topic, and facts. Different from standard language understanding tasks, Chalamalasetti et al. (2023) presents a novel method to evaluate the ability of LLMs to act as '*situational*' language understanding agents (Schlangen 2023). They do so by assigning roles to LLMs and generate dialogues resembling word games such as taboo, and test the language generating and instruction following abilities of LLMs based on the quality of game-play leading to successful target word prediction.

Although we propose a similar approach to evaluation by utilising conversations of such a word game, our work differs from theirs in two ways: (a) they use LLM-generated conversations while we rely on an existing dataset of conversations; (b) they do not employ dialects in their conversations while the dataset we use contains information about the dialects of the human speakers.

6 Conclusion

Although superlative performances have been reported on LLMs in recent times, recent work shows the performance gap between US English and other dialects of English. Our paper presents a first-of-its-kind evaluation of the dialect robustness of LLMs using their ability to predict target words in game-playing conversations. We use a dataset of target-word-masked conversations between US English speakers and those between Indian English speakers playing a game of taboo. We evaluate pre-trained and fine-tuned versions of one open-source and two closed-source models, on two tasks: target word prediction (TWP) and target word selection (TWS). Our results show that the LLMs indeed perform better for en-US as compared to en-IN on both tasks, with the average performance being higher by 12.66 and 17.4 on similarity and accuracy scores across all configurations. This shows that LLMs marginalise or discriminate against speakers of the Indian dialect. We also observe that pre-trained models report a degraded performance on conversations created using both rule-based (en-MV) and LLM-based (en-TR) transformations, as compared to their source conversations (en-US and en-IN respectively). However, fine-tuning on en-MV yields a greater improvement in the task performances, as compared to that on en-TR. This shows that the transformations that introduce dialectal information about a national variety help in improving the dialect robustness of LLMs more than the transformations that remove the said dialectal information. Finally, our error analysis demonstrates that, while most errors are mitigated by providing options for masked target word (TWS; in both pre-trained and fine-tuned variants), LLMs struggle to interpret target words based on the shared cultural context between speakers.

Our extension M-MD3 is a dataset for TWP and TWS based on MD3, consisting of four subsets: en-US, en-IN, en-MV, and en-TR. The dataset opens opportunities for future evaluations of dialect robustness using similar conversation-based tasks. Our evaluation methodology can also be applied to other existing dialogue and discourse datasets, to evaluate the ability of LLMs on properties other than dialect robustness.

Limitations

The original MD3 paper states that their dataset may over-represent Western entities to some degree. Therefore, it is possible that Indian speakers faced difficulties with the terms. Having said that, the instances selected for our dataset are the ones where the Indian players guessed the word correctly. We have not performed a detailed qualitative analysis of these conversations, except for a cursory sanity check. We also assume that the dialect of English from each locale is homogeneous. Assuming that en-IN is the English spoken in every region of India is an unrealistic generalization of the diversity of dialects of English. In terms of model fine-tuning, our paper also does not cover the impact of quantization and different fine-tuning (including cross-dialect) techniques on the task.

Ethics Statement

We use a publicly available dataset of conversations consisting of human players engaged in a game of taboo. The topics discussed in the dataset are fairly general, and are unlikely to cause distress. The error analysis was performed by one of the authors of the paper. The AI-transformed (en-TR) conversations may contain biased output, arising due to inherent properties of GPT-based models.

References

Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476. Online: Association for Computational Linguistics.

Bolton, K. 2012. World Englishes and linguistic landscapes. *World Englishes*, 31(1): 30–33.

Chalamalasetti, K.; Götze, J.; Hakimov, S.; Madureira, B.; Sadler, P.; and Schlangen, D. 2023. clembench: Using Game Play to Evaluate Chat-Optimized Language Models as Conversational Agents. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 11174–11219. Singapore: Association for Computational Linguistics.

Chen, Y.; Liu, Y.; Chen, L.; and Zhang, Y. 2021. DialogSum: A Real-Life Scenario Dialogue Summarization Dataset. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 5062–5074. Online: Association for Computational Linguistics.

Chen, Z.; Chen, L.; Chen, B.; Qin, L.; Liu, Y.; Zhu, S.; Lou, J.-G.; and Yu, K. 2022. UniDU: Towards A Unified Generative Dialogue Understanding Framework. In Lemon, O.; Hakkani-Tur, D.; Li, J. J.; Ashrafzadeh, A.; Garcia, D. H.; Alikhani, M.; Vandyke, D.; and Dušek, O., eds., *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 442–455. Edinburgh, UK: Association for Computational Linguistics.

Dettmers, T.; Lewis, M.; Shleifer, S.; and Zettlemoyer, L. 2022. 8-bit Optimizers via Block-wise Quantization. arXiv:2110.02861.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. arXiv:2305.14314.

Dey, S.; and Desarkar, M. S. 2023. Dial-M: A Masking-based Framework for Dialogue Evaluation. In Stoyanchev, S.; Joty, S.; Schlangen, D.; Dusek, O.; Kennington, C.; and Alikhani, M., eds., *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 77–84. Prague, Czechia: Association for Computational Linguistics.

Eisenstein, J.; Prabhakaran, V.; Rivera, C.; Demszky, D.; and Sharma, D. 2023. MD3: The Multi-Dialect Dataset of Dialogues. arXiv:2305.11355.

Faisal, F.; Ahia, O.; Srivastava, A.; Ahuja, K.; Chiang, D.; Tsvetkov, Y.; and Anastasopoulos, A. 2024. DIALECTBENCH: A NLP Benchmark for Dialects, Varieties, and Closely-Related Languages. In *Proceedings of the 2024 Association for Computational Linguistics (ACL)*.

Gliwa, B.; Mochol, I.; Biesek, M.; and Wawer, A. 2019. SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization. In Wang, L.; Cheung, J. C. K.; Carenini, G.; and Liu, F., eds., *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, 70–79. Hong Kong, China: Association for Computational Linguistics.

Held, W.; Ziems, C.; and Yang, D. 2023. TADA : Task Agnostic Dialect Adapters for English. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 813–824. Toronto, Canada: Association for Computational Linguistics.

Joshi, A.; Dabre, R.; Kanojia, D.; Li, Z.; Zhan, H.; Haffari, G.; and Dippold, D. 2024. Natural Language Processing for Dialects of a Language: A Survey. arXiv:2401.05632.

Khandelwal, K.; Tonneau, M.; Bean, A. M.; Kirk, H. R.; and Hale, S. A. 2024. Indian-BhED: A Dataset for Measuring India-Centric Biases in Large Language Models. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*, GoodIT '24, 231–239. New York, NY, USA: Association for Computing Machinery. ISBN 9798400710940.

Liu, Y.; Held, W.; and Yang, D. 2023. DADA: Dialect Adaptation via Dynamic Aggregation of Linguistic Rules. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 13776–13793. Singapore: Association for Computational Linguistics.

Llama Team. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Negro, S. D.; and Vietti, A. 2006. The interplay of dialect and the standard in anonymous street dialogues: Patterns of variation in northern Italy. *Language Variation and Change*, 18(2): 179–192.

OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155.

Qin, L.; Gupta, A.; Upadhyay, S.; He, L.; Choi, Y.; and Faruqui, M. 2021. TIMEDIAL: Temporal Commonsense Reasoning in Dialog. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 7066–7076. Online: Association for Computational Linguistics.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv:1908.10084.

Ryan, M. J.; Held, W.; and Yang, D. 2024. Unintended Impacts of LLM Alignment on Global Representation. arXiv:2402.15018.

Sai, A. B.; Mohankumar, A. K.; Arora, S.; and Khapra, M. M. 2020. Improving Dialog Evaluation with a Multi-reference Adversarial Dataset and Large Scale Pretraining. *Transactions of the Association for Computational Linguistics*, 8: 810–827.

Schlangen, D. 2023. On General Language Understanding. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 8818–8825. Singapore: Association for Computational Linguistics.

Sun, K.; Yu, D.; Chen, J.; Yu, D.; Choi, Y.; and Cardie, C. 2019. DREAM: A Challenge Data Set and Models for Dialogue-Based Reading Comprehension. *Transactions of the Association for Computational Linguistics*, 7: 217–231.

Ueyama, A.; and Kano, Y. 2023. Dialogue Response Generation Using Completion of Omitted Predicate Arguments Based on Zero Anaphora Resolution. In Stoyanchev, S.; Joty, S.; Schlangen, D.; Dusek, O.; Kennington, C.; and Alikhani, M., eds., *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 282–296. Prague, Czechia: Association for Computational Linguistics.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13484–13508. Toronto, Canada: Association for Computational Linguistics.

Wikipedia. 2023. Taboo (game).

Xiao, Z.; Held, W.; Liu, Y.; and Yang, D. 2023. Task-Agnostic Low-Rank Adapters for Unseen English Dialects. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 7857–7870. Singapore: Association for Computational Linguistics.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; Du, Y.; Yang, C.; Chen, Y.; Chen, Z.; Jiang, J.; Ren, R.; Li, Y.; Tang, X.; Liu, Z.; Liu, P.; Nie, J.-Y.; and Wen, J.-R. 2023. A Survey of Large Language Models. arXiv:2303.18223.

Ziems, C.; Held, W.; Yang, J.; Dhamala, J.; Gupta, R.; and Yang, D. 2023. Multi-VALUE: A Framework for Cross-Dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics.

A Dataset Construction

Table 5 describes the example conversations from extended MD3 and their corresponding masked versions from M-MD3. We mask the turn where the guesser utters the target word to help with formulating our downstream tasks. We mask the target word by finding the exact match in the conversation as shown in the conversations from Table 5. In case of conversations where an exact match is not found (such as *planets*), we find the utterance that is most similar to the target word using the similarity score[10]. The rest of the conversation is then pruned to make the masked target word (represented by '[MASK]') the last token in the conversation.

B Transformation Issues

We transform conversations from en-IN into en-TR, by prompting[11] GPT-4 to remove exaggerations and dialectal information. As mentioned in Table 6, a *'typical'* transformed conversation maintains the semantic meaning but only differs from the original conversation grammatically. A *'bad'* example deviates largely from the expected output, and such examples are excluded from the final set of conversations used in our evaluation.

C Errors

Table 7 describes additional examples for all identified error types[12]. As mentioned, each conversation can be classified under multiple error types. For example, the conversation about the target word–*'Ryan Reynolds'* is classified as **CC**, but can also be classified as **PF**.

[10] Described in the *Metrics* section of the main paper.

[11] The exact prompt can be found in the *Methodology* section of the main paper.

[12] defined in the *Error Analysis* section of the main paper.

Target Word	**en-IN**	**Masked en-IN**

Fisherman	`Describer: Uh. What do you call if we, what will be there in the water?`
	`Guesser: Fishes`
	`Describer: Who will catch that?`
	Guesser: ***Fisherman***.
		`Describer: Uh. What do you call if we, what will be there in the water?`
		`Guesser: Fishes`
		`Describer: Who will catch that?`
		Guesser: ***[MASK]***

Target Word	**en-US**	**Masked en-US**
Planet	`Describer: These are hard words. um Okay. So there’s. the Sun and the Moon and all the rest of them.` Guesser: And all the ***planet***s? **(**`Describer: Yes.`**)**	`Describer: These are hard words. um Okay. So there’s. the Sun and the Moon and all the rest of them.` Guesser: ***[MASK]***

Table 5: Masking conversations from the extended MD3 to create M-MD3. The text such as ***this*** represents the target word utterance by the guesser which is masked (represented by, ***[MASK]*** in the M-MD3 version of the conversation. The rest of the original conversation is pruned as represented text in parentheses.

Type	en-IN	en-TR
Typical	Describer: **(**Uh**)**. What do you call ***if we, what will be there*** in the water? Guesser: Fish**(**es**)** Describer: Who ***will catch that***? Guesser: Fisherm***a***n.	Describer: **(∅)** What do you call ***the creatures*** in the water? Guesser: Fish**(∅)**. Describer: Who ***catches them***? Guesser: Fisherm***e***n.
Bad	Describer: There. is a. there is a character in a movie Guesser: um Describer: It’s a very famous movie and it’s a very. where is a where you can see famous dialogue called I am still gorgeous Guesser: uh. ok. uh Describer: character name. compare like Marvel movie Guesser: So. uh Describer: very muscular body gives scientifically. injections Guesser: ok Describer: His body will grow Guesser: uh Captain America	The character being described is from a well-known movie and is known for the line “I am still gorgeous.” This character is similar to those found in Marvel movies and has a very muscular physique, which is achieved through scientific injections that cause his body to grow. The character being guessed is Captain America.

Table 6: Example *transformations* of en-IN to en-TR. We utilise GPT-4 to generate the response. The text in parentheses refers to the omission/removal of certain filler and exaggerated words, and the text such as ***this***, refers to the words or sentences that were rephrased to convey the original meaning.

Type	Target Word	Conversation	Model Prediction
AD	Fisherman	Describer: Okay. Okay. A. guy um wants to um okay. Guy catching something in the water. Guesser: [MASK]	Fish
WD	Atlantic Ocean	Describer: One of the. of the of world. Guesser: Of the seventh wonder of the world. Taj mahal? Is it regarding sea? Describer: No no no the. Towards the bottom of India. Guesser: Is it regarding Describer: what we have? Guesser: [MASK]	Kanyakumari
BDD	Russian Language	Describer: Ok. Ah Largest continent in the world Guesser: Ok. Describer: Ah like area wise. Which country? Guesser: Largest. vast area. vast area? Russia but. Describer: We need to add N over there at the end. Guesser: Russian Describer: We speak Guesser: What they speak? Describer: Yeah. Ok. Guesser: [MASK]	Russian
CC	Ryan Reynolds	Describer: It is like. One of the. Pen name. which we used in school school days. Guesser: Cello point pen. Fine Grip Describer: No no no Guesser: Reynolds Describer: Uh yeah yeah Guesser: This is a second word or first word. Describer: Yeah this is second word Guesser: First word is. Name Describer:Yeah name related to the same Guesser: [MASK]	Flair
PF	Steve Jobs	Describer: Ok. He is a famous person and he is a. a. for. what we call? . Um now it is a. Its. giving competition to Android. what we call? Guesser: ok. so he is the fond ok sorry Describer: he is a founder of so and so company. Its a U. S. company Guesser: so it is giving competition to Android means Google ok.. So Describer: and he is the founder of that company Guesser: [MASK]	Steve
ERR	Podium	Describer: Okay um. uh. well I isn't sure I'm not sure but uh letting are seeing. Well it's like preacher are churching. I am standing behind this. uh. in in used for speaker. Guesser: [MASK]	Pulpit

Table 7: Example conversations (*'Conversation'*) for each error type (*'Type'*) along with the reference target word (*'Target Word'*) and the generated target word (*'Model Prediction'*). All model predictions are generated using the pre-trained variants of GPT-4 and LLAMA-3.