

Examining the Efficacy of LLMs in Comprehending Homophones

Anonymous ACL submission

Abstract

The idea of this work is inspired by censorship in social media. Social media users use manually created homophone words to bypass censorship. In this paper, we built testsets of manually created homophone words in different languages, and evaluate LLMs ability to comprehend sentences with such homophones.

Few work focus on homophone words in text, and no previous study was focusing on manually created homophone words. This paper is the first one to study LLMs' ability in recognizing manually created homophones in dialogue.

1 Introduction

Homophone words are words that spells different, means different and sounds similar. For example, the word 'night' and 'knight' are a pair of homophone words. The phenomenon of homophone is rare in English, however, for certain language, like Chinese, homophones are very frequent and can be easily created.

Censorship is common in social media, and social media users sometimes manually create homophone words to bypass censorship. As an example, if the word 'control' is banned, users can instead use a word like 'kentroll', a word sounds as similar as possible to 'control' and at the same time spells as different as possible, to bypass censorship.

The idea of this paper is inspired by such clever trick social media uses take advantage of. In this paper, we built 3 testset, in Chinese, English and Spanish, respectively. Each testset include 100 examples, and an example includes a sentence, a word, an answer, and records of the correctness of some LLMs' answer to the homophone recognition question generated from this example.

We run the testset on GPT-3.5, GPT-4 classic, Ernie-3.5 and Ernie-4.0. We discovered that GPT-4 performs the best on all 3 testsets. The 2 Chinese

built models, Ernie-3.5 and Ernie-4.0, have similar performance, with Ernie-4.0 performs slightly better than Ernie-3.5. Both Chinese built LLMs perform better than GPT-3.5 for Chinese testset. However for Spanish testset, the 2 Chinese built models perform extremely bad. GPT-3.5 performs ok on English testset and Spanish testset, but performs poorly on Chinese testset.

The testsets are created in .csv format. All testsets and related files can be found on Github [1].

2 Related Works

We didn't find much helpful related works. Most papers in this area focus on recognizing homophone words in voice chat, other than text chat.

3 Baseline

We measure LLMs' ability to comprehend homophones by accuracy of question-answering. The questions are carefully designed so that there's no direct clue of correct answer in question. The baseline is randomly guess a word. There are about 30,000 commonly used English words, so a random guess has accuracy $\frac{1}{30,000}$, which is close enough to zero. The logic of Chinese and Spanish testsets are similar. So, any accuracy above zero means that the LLM have some understanding of these homophone questions.

4 Data Usage

This work does not rely on any external database. All the data in testsets and code in scripts are written by co-author Clair Ross and Nuoqi Liu.

5 Contribution

In this work we discover a new problem to solve, that is comprehending manually created homophone words in text dialogue. We proposed a method to measure LLM's ability to understand homophones. And we created a testset for such

measurement. We are the first to study the performance of a few state-of-the-art LLMs in a few languages. Also we propose some known issues in our methods, shed light on future study in this area.

6 Details of Implementation

6.1 Format of testset

In this work we implement 3 testsets, in Chinese, English and Spanish respectively. An example row in English testset includes a sentence, a word and an answer. Table 1 shows a few examples.

Table 1: Examples of testset_english.csv

sentence	word	answer
The drywer is drunk	drywer	driver
I want a marshrum burger	marshrum	mushroom
Cat is one kind of maimol	maimol	mammal
The san is so shiny today	san	sun
She has furtell imagination	furtell	fertile

In each row, one word in the sentence makes little sense literally and is actually a manually created homophone. The column 'word' indicates explicitly which word is homophone. The column 'answer' gives the original word of the manually created homophone.

For Chinese and Spanish testsets, the format is a bit different. A new column 'answer_english' is added, indicating the English translation of correct answer. This is for the convenience of non-native Chinese and Spanish speaker to easily use the testset.

6.2 Creation of Testset

All 3 testsets are created manually. Each testset includes 100 human written sentences and human created homophones. The homophones are aimed to spell as different as possible from original word, and at the same time sounds as similar as possible.

6.3 Question Generation

For each example in one testset, a corresponding question needs to be generated, as the input of LLM chatbot. The question is asked in English, regardless of which testset it is generated from. Take Spanish testset as an example. Suppose the sentence is

Quiero ir ha todos los festivales estacionales con mis amigos este verano

And the homophone word is "ha", the question will goes as

In the sentence 'Quiero ir ha todos los festivales estacionales con mis amigos este verano', the word 'ha' is a homophone word, can you tell its true meaning?

In the Github repository of this project, a python script "input_generation.py" implements this procedure.

6.4 Running the Test

The test goes in a zero-shot manner. The question is the only input, no extra information or rules explanation is presented to the model. Each question answering pair happens in a new session. The correctness of an answer is judged solely by if it contains the correct answer. For English testset, it means a correct should contain the original word. For Chinese and Spanish testsets, an answer is marked as correct by containing either the original word in English or Chinese & Spanish, respectively.

7 Results

7.1 Final Results

The accuracy of each model on each testset is shown in Table 2:

Table 2: Accuracy of LLMs' answers on testsets

	Chinese	English	Spanish
GPT-3.5	28%±4.5%	63%±4.8%	63%±4.8%
GPT-4	60%±4.9%	90%±3.0%	82%±3.8%
Ernie-3.5	50%±5.0%	62%±4.9%	6%+?%
Ernie-4.0	63%±4.8%	83%±3.8%	10%+?%

7.2 Comments on Results

Accuracies are calculated based on number of correct answers out of 100 questions, so all the results are integers. Ernie-3.5 and Ernie-4.0 crashes when the input contains too much Spanish, so they got very low accuracy in that scenario.

GPT-4 generally performs the best in all scenario. Ernie bots are developed by a Chinese company, and they performs fine on Chinese testset, as expected. Chinese questions are generally more difficult than English and Spanish questions, for that there's no clue in the spell for Chinese homophone words, which coincide with the results that

all models tested got lower accuracy on Chinese testset. GPT-4 outperforms GPT-3.5 by a lot, and Ernie-4.0 only outperform Ernie-3.5 by a little bit.

8 Error Analysis

Accuracy is measured by numbers of correct answer out of 100 questions, thus it follows binomial statistics distribution. For binomial statistics distribution, the average value is p and the standard deviation is $\sqrt{np(1-p)}$. Estimation of error shown in table 2 is calculated based on this formula.

For Ernie-3.5 and Ernie-4.0 in Spanish, however, previous error estimation technique is not applicable. Ernie chatbot is overall weak on Spanish and crashes a lot when long series of Spanish text is inputted. As a result of crash, Ernie bots got extremely low accuracy on Spanish testset. It is hard to tell Ernie's true ability to comprehend Spanish homophone, still we believe the true accuracy shall be some value much higher than those shown in table 2.

9 Other Findings

Although methods to improve LLM's ability to comprehend homophones is not the main focus of this paper, we did find a few simple methods that may help.

9.1 Forced Answer

The most common type of error LLM makes is refusal. It claims that there's no enough context or information, so it cannot give an answer. However, simply ask the same question again, and it may gives a correct answer. In the below example, it takes twice asking that GPT-3.5 gives a correct answer.

Me: In the sentence 'I want a marshrum burger', the word 'marshrum' is a homophone word, can you tell its true meaning?

GPT-3.5: It seems like "marshrum" might be a misspelling or a made-up word. Without any context or known definition, it's hard to determine its true meaning. If it's a homophone for another word, it might be a creative way of saying something else. Can you provide any additional context or clarify if it's a term you've encountered elsewhere?

Me: In the sentence 'I want a marshrum burger', the word 'marshrum' is a homophone word, can you tell its true meaning?

GPT-3.5: Ah, I see what you're getting at! In this case, "marshrum" is likely a playful or intentional misspelling of "mushroom." A mushroom burger would typically contain mushrooms as one of its main ingredients. So, in this context, "marshrum" is a homophone for "mushroom," used humorously or creatively.

Of course, sometimes this simple trick will not work. Also, sometimes it gives a wrong answer after repeated requests.

10 Known Issues

10.1 Clues from the context

Nothing prevents the model to answer these questions based only on context, but not on pronounce of homophone words. Sometimes the model answers explicitly that it find the correct answer based on context. Also sometimes the model declares how it reasons to find an answer, and the way it shows has nothing to do with sounds and pronounces.

10.2 Clues from the spell

For English and Spanish, it is almost impossible to manually create a homophone word that spells totally different and sounds similar to the original word. Sometimes the edit distance between homophone word and original word is only 2 or 3, and nothing prevent LLM to find original word based only on edit distance based typo correction. Sometimes LLM answer explicitly that it find the answer by examing the spell.

As a comparison, for Chinese words, it is very easy to find an another word that differ in every letter and sounds exactly the same. For the Chinese testset, all the homophone words are carefully created in this style, so there's no hint in the spell at all. And we did see a clear drop of accuracy, when a same model comes from English or Spanish to Chinese testset. An exception is Ernie, it is overall weak in Spanish language, and the accuracy is higher for Chinese, but that is another story.

10.3 Multiple Answers

Another issue is that no measurement is there to stop model from guessing many answers. Although

in practice, usually the model will not guess a lot of answer. In most cases, it gives either one answer or no answer at all. In some rare cases (about 5% chance) it will gives 2 or 3 answers, one of them can be correct, and the other answers are ridiculously wrong.

10.4 Overall weak Model

There's another issue when testing both Ernie-3.5 and Ernie-4.0 on Spanish questions. Ernie bots are overall weak in Spanish. It cannot handle long Spanish inputs and crash a lot. The extremely low accuracy for Ernie on Spanish is mostly due to such crash. The actually ability for Ernie series bot to comprehend homophones is expected to be higher than reflected by the accuracy in this paper.

11 Conclusion

In this paper, we proposed the concept of manually created homophones and point out the importance for LLM to recognize such expression. We defined a way to measure a LLM's ability to comprehend homophone expressions in dialogue. We build 3 testsets in Chinese, English, Spanish, respectively. We find out that GPT-4 performs well in all 3 languages, GPT-3.5 performs ok in English and Spanish but poorly in Chinese. Ernie-4.0 outperform Ernie-3.5 by a little bit, both performs ok in Chinese and English, but extremely poorly in Spanish. We realize some issue in our method, including hint in spell and context, and low accuracy due to overall weak performance.

Acknowledgments

We appreciate professor Vered Shwartz's guidance and University of British Columbia's excellent academic environment.

custom

References

- [1] Clair Ross & Nuoqi Liu. *CPSC-532V project*. 2024. URL: <https://github.com/LNQ1997/CPSC532V-project>.

A Appendix - Full Testset Tables

There's bug displaying Chinese and large tables in this environment, so we post tables in images. Csv format of testsets can be found in Github Repository [1]

	sentence	word	answer_english	answer_chinese	GPT-3.5	GPT-4	Ernie-3.5	Ernie-4.0
1	这是我新买的见磐	见磐	keyboard	键盘	Y	Y	Y	Y
2	我早餐吃了面包和机但	机但	egg	鸡蛋	Y	Y	Y	Y
3	希望站征早日结束	站征	war	战争	N	Y	Y	Y
4	考试期间禁止使用首姬	首姬	cellphone	手机	N	Y	Y	N
5	他经常七负同学	七负	bully	欺负	N	Y	Y	N
6	我要买一个牛肉翰宝	翰宝	hamburger	汉堡	N	Y	N	N
7	请注意观察宏律登	宏律登	traffic light	红绿灯	N	N	N	N
8	桌上有一支刚比	刚比	pen	钢笔	N	Y	Y	Y
9	我想吃由闸鸡翅	由闸	fried	油炸	N	Y	N	Y
10	锤子标志象征着攻仁	攻仁	laborer	工人	N	N	N	Y
11	要求总理磁直的呼声越来越高	磁直	resign	辞职	N	Y	Y	N
12	俄罗斯受到了西方的至才	至才	sanction	制裁	N	Y	N	N
13	政府对今年的经济预期十分勒关	勒关	optimistic	乐观	N	Y	Y	Y
14	在公园中央有一个貂项	貂项	statue	雕像	N	Y	N	Y
15	朝鲜政府改变了对待韩国的正测	正测	policy	政策	N	Y	Y	Y
16	议长宣布担任林石总统	林石	temporary	临时	N	N	N	N
17	美国停止了对高棉共和国的缘柱	缘柱	assistance	援助	N	Y	Y	Y
18	数千微信用户呼吁蟪晓聋辞职	蟪晓聋	Zhang Xiaolong	张小龙	N	N	Y	N
19	学生们普遍对笑张意见很大	笑张	headmaster	校长	N	N	N	N
20	这些仕已经过期的食物	仕	are	是	N	Y	N	Y
21	斐基发动机的制造仍然是个难题	斐基	airplane	飞机	N	Y	N	N
22	今日外长和伊朗总统通了垫画	垫画	phone call	电话	Y	Y	Y	N
23	当地樯仕目前十分敏感	樯仕	situation	局势	N	Y	Y	N
24	污磕蓝宣布开始新一轮征兵	污磕蓝	Ukraine	乌克兰	N	Y	N	Y
25	工人们经常抱怨架七太少	架七	holiday	假期	N	N	N	Y
26	当地群众都对判决忆份甜鹰	忆份甜鹰	righteous indignation	义愤填膺	N	N	Y	Y
27	此案已经成为刑侦界的经典暗立	暗立	case study	案例	Y	Y	Y	Y
28	两天后案情终于有了重大近斩	近斩	progress	进展	Y	Y	Y	Y
29	她学习十分弩例	弩例	hard working	努力	Y	Y	Y	Y
30	案件侦破的棉渡极大	棉渡	level of difficulty	难度	N	Y	N	Y
31	视野中的画面越来越魔呼	魔呼	vague	模糊	N	Y	Y	Y
32	他出生于一个缚桂人家	缚桂	rich	富贵	N	Y	Y	Y
33	小学时他就经常矿氩	矿氩	skip class	旷课	N	Y	N	N
34	符合这些调箭的人很多	调箭	requirement	条件	N	Y	N	N
35	那半截香烟成了破案的观剑	观剑	key	关键	N	Y	Y	Y
36	老虎行动时几乎不会发出升阴	升阴	sound	声音	N	Y	Y	Y
37	他出发那一天是腥七武	腥七武	Friday	星期五	Y	Y	Y	Y
38	一家人决定去野生动物园字架油	字架油	road trip	自驾游	N	Y	N	Y
39	铨续的话让陈老头怒不可遏	铨续	son-in-law	女婿	N	N	Y	N
40	他因生活节俭而被怀疑是供铲档	供铲档	communist	共产党	Y	Y	N	N
41	一名爱好者被卡在了洞穴钟兼	钟兼	middle	中间	N	N	N	Y
42								

Figure 1: Chinese Testset 01

40	铨续的话让陈老头怒不可遏	铨续	son-in-law	女婿	N	N	Y	N
41	他因生活节俭而被怀疑是供铲档	供铲档	communist	共产党	Y	Y	N	N
42	一名爱好者被卡在了洞穴钟兼	钟兼	middle	中间	N	N	N	Y
43	她是来自牛津大学的育裁声	育裁声	talented student	高材生	Y	Y	N	Y
44	两人早就成为了英国洞穴协会的惠缘	惠缘	member	会员	N	Y	Y	Y
45	两人每次都能话薛围疑	话薛围疑	turn danger into safety	化险为夷	N	N	N	N
46	曾有无数探险家来这个洞穴追求次基	次基	excitement	刺激	N	N	N	N
47	莫斯带着队友记絮前进	记絮	continue	继续	N	N	N	Y
48	爬上斜坡后可以看到前往纪念堂的稻露	稻露	road	道路	N	Y	Y	Y
49	这些规则同样适用于洞穴探险等集线运动	集线	extreme	极限	Y	N	Y	Y
50	病人来时呼吸十分籍醋	籍醋	hasty	急促	N	Y	N	Y
51	病人的雪鸭正常	雪鸭	blood pressure	血压	N	Y	N	Y
52	这位衣声平日工作非常认真	衣声	doctor	医生	N	N	Y	Y
53	两个星期之后,枕段结果公布	枕段	diagnosis	诊断	N	N	N	Y
54	医生们决定立刻让病人溪养	溪养	oxygen inhalation	吸氧	N	N	N	N
55	她得知男朋友有比较严重的费研	费研	pneumonia	肺炎	Y	Y	Y	Y
56	此时病人已经十分需若	需若	weak	虚弱	N	Y	Y	Y
57	看着诊断报告,他感到十分厚毁	厚毁	regret	后悔	Y	N	Y	Y
58	老虎的习性十分胸蚕	胸蚕	cruel	凶残	N	Y	N	N
59	旧缘人员在十分钟之后赶到	旧缘	rescue	救援	N	Y	N	Y
60	这场惨剧的音颖至今仍未完全消除	音颖	shadow	阴影	N	Y	N	N
61	这支部队已经严重缺少但要	但要	ammunition	弹药	N	Y	N	Y
62	她的手臂上留下了明显的商巴	商巴	scar	伤疤	N	Y	Y	Y
63	苏联秘密同意对日本轩站	轩站	declare war	宣战	N	Y	N	Y
64	这样做能得到什么郝鑫呢	郝鑫	benefit	好处	N	Y	N	Y
65	二人商量如何刮芬所得的利润	刮芬	divide up	瓜分	N	Y	N	Y
66	这条铁路全长三千工李	工李	kilometer	公里	Y	Y	Y	Y
67	这座城市东纪平均气温为零度	东纪	winter	冬季	N	Y	N	N
68	三千名站芙负责修建这条铁路	站芙	prisoner of war	战俘	N	N	N	N
69	战俘们被强迫参与牢洞	牢洞	labor	劳动	N	Y	N	N
70	西伯利亚就是一座天然间欲	间欲	prison	监狱	N	N	Y	N
71	二战受害者对酥莲此举喜闻乐见	酥莲	Soviet Union	苏联	N	Y	N	N
72	一部分人负责种植钭逗	钭逗	potato	土豆	N	N	N	Y
73	此时这个国家摆费带星	摆费带星	everything remains to be done	百废待兴	N	Y	Y	Y
74	小队花了两个月收集足够的速财	速财	material	素材	N	Y	Y	N
75	就连看守的日常尹石也十分匮乏	尹石	diet	饮食	N	N	N	N
76	他入狱的第一餐,竟然是蛮投	蛮投	bun	馒头	N	N	Y	Y
77	战争让所有人认识到了攻液的重要性	攻液	industry	工业	N	N	N	N
78	但愿西伯利亚的苦难永远成为立矢	立矢	history	历史	N	Y	Y	N
79	她的身上盖着两条眠背	眠背	quilt	棉被	Y	Y	Y	Y
80	案发现场发现了项链、育根邪等物品	育根邪	high heels	高跟鞋	N	Y	N	N
81	几名民警言揪了一番,但没有得出结论	言揪	study	研究	N	Y	Y	Y

Figure 2: Chinese Testset 02

80	案发现场发现了项链、膏根邪等物品	膏根邪	high heels	高跟鞋	N	Y	N	N
81	几名民警言揪了一番，但没有得出结论	言揪	study	研究	N	Y	Y	Y
82	这个凶手简直就是遍钛	遍钛	pervert	变态	N	N	Y	Y
83	汤老师平时从来不呵久	呵久	drink alcohol	喝酒	N	N	N	N
84	此案仍然有很多疑点需要姐市	姐市	explain	解释	N	Y	Y	Y
85	此案让汤老师的铭升被彻底败坏	铭升	reputation	名声	N	Y	N	Y
86	队长在会议上提出了两点忆箭	忆箭	suggestion	意见	Y	Y	N	Y
87	窃贼迅速从疮护进入了二楼阳台	疮护	window	窗户	Y	Y	Y	Y
88	这是一辆从北京开往审羊的火车	审羊	Shenyang	沈阳	N	N	N	Y
89	此人是此案的第一闲移仁	闲移仁	suspect	嫌疑人	N	N	N	Y
90	男人的情绪立刻变得蝎丝抵礼	蝎丝抵礼	hysterical	歇斯底里	N	N	Y	Y
91	他是一个极端字斯的人	字斯	selfish	自私	N	Y	Y	Y
92	这一次他更加从容地玩弄着裂悟	裂悟	prey	猎物	N	N	N	N
93	他的童年生活并不性芙	性芙	well-being	幸福	N	Y	Y	Y
94	最近小区内的范最事件明显增多	范最	crime	犯罪	N	Y	Y	Y
95	他要求手下必须勇敢以及中城	中城	loyalty	忠诚	Y	Y	Y	N
96	四名站市被派遣驻守此地	站市	soldier	战士	Y	Y	Y	Y
97	他十分热衷于这种疯狂的竟塞	竟塞	competition	竞赛	Y	Y	Y	Y
98	幸好事后店主表示笋师并不大	笋师	loss	损失	Y	Y	Y	N
99	专家们仔细调查了当地的狙旻	狙旻	resident	居民	N	N	N	N
100	向东大约600米还有一个倒与	倒与	island	岛屿	N	N	N	Y
101	这个肿卒仍处于石器时代的水平	肿卒	race	种族	N	N	N	N

Figure 3: Chinese Testset 03

	sentence	word	answer	GPT-3.5	GPT-4	Ernie-3.5	Ernie-4.0
2	The drywer is drunk	drywer	driver	Y	Y	Y	N
3	Next Monday is a national howllydey	howllydey	holiday	Y	Y	Y	Y
4	The san is so shiny today	san	sun	Y	Y	N	Y
5	I want a marshrum burger	marshrum	mushroom	N	Y	Y	N
6	Intruders have destroyed the housepitle	housepitle	hospital	N	Y	N	Y
7	Cat is one kind of maimol	maimol	mammal	N	Y	Y	Y
8	This topic is trending on sowshow media	sowshow	social	Y	Y	N	Y
9	Refugees seek shallter at school	shallter	shelter	Y	Y	Y	Y
10	There's no sign of brickthrow for ceasefire talk	brickthrow	breakthrough	Y	N	N	Y
11	He filled in a sirway about the protest	sirway	survey	Y	Y	Y	Y
12	The tryel is expected to last for 6 month	tryel	trial	Y	Y	Y	Y
13	He worries about possible whalelense in the street	whalelense	violence	N	Y	N	N
14	Two passengers are injured in the bus crash aksident	aksident	accident	Y	Y	Y	Y
15	An eight year old girl is the sole serwaiwor	serwaiwor	survivor	N	Y	N	Y
16	South Africa is one of the most diwylloped country in Africa	diwylloped	developed	N	Y	Y	Y
17	The driver lost kentroll of the car	kentroll	control	Y	Y	Y	Y
18	The boat caught fire after inpackt on a rock	inpackt	impact	Y	Y	Y	Y
19	She has furtell imagination	furtell	fertile	Y	Y	N	Y
20	Flood has caused synnyfikent damage to this area	synnyfikent	significant	Y	Y	Y	Y
21	Xia is the first dinicety of China	dinicety	dynasty	N	Y	Y	Y
22	News media usually describe him as a diktiter	diktiter	dictator	Y	Y	Y	Y
23	Such naïve perspective shall only exist in kids' kognashion	kognashion	cognition	Y	Y	Y	Y
24	Central control cistern is down for 2 hours	cistern	system	Y	Y	Y	Y
25	China become a superpower in only a few deckiz	daikiz	decades	Y	Y	Y	Y
26	They suspect he has done something creamanol	creamanol	criminal	N	Y	N	Y
27	Drug trifaking is banned here	trifaking	trafficking	N	Y	Y	Y
28	You can get a callpen for purchasing this item	callpen	coupon	N	N	N	Y
29	My dad bout a land 20 years ago	bout	bought	Y	Y	Y	Y
30	He tried to save every pani in his life	pani	penny	Y	Y	N	Y
31	They've been here the last wiik	wiik	week	Y	Y	Y	Y
32	I need to attend a bezniss meeting in Guangzhou	bezniss	business	Y	Y	Y	Y
33	The campaign lasted for one manse	manse	month	Y	Y	Y	Y
34	Another foto of him is revealed later	foto	photo	Y	Y	Y	Y
35	I've made a serious missdack	missdack	mistake	Y	Y	Y	Y
36	He think of seavirel methods	seavirel	several	Y	Y	N	Y
37	People wonder what is the last city he will kenkour	kenkour	conquer	N	N	N	Y
38	She once had a happy trip to Yijept	Yijept	Egypt	N	Y	N	Y
39	She spend a whole weekend on reading metireal	metireal	material	Y	Y	Y	Y
40	The company received a lot of rizeme this weak	rizeme	resume	N	Y	N	N
41	There are a lot of blike people living in this city	blike	black	N	N	N	N
42	She wood rather find a new assistant	wood	would	Y	Y	Y	Y

Figure 4: English Testset 01

40	The company received a lot of rizeme this weak	rizeme	resume	N	Y	N	N
41	There are a lot of blike people living in this city	blike	black	N	N	N	N
42	She wood rather find a new assistant	wood	would	Y	Y	Y	Y
43	That assistant is experienced in working for selebraty	selebraty	celebrity	Y	Y	Y	Y
44	She doesn't want to mention any ixbeerinse about that day	ixbeerinse	experience	Y	Y	Y	N
45	She discovered that her idol is nothing like what she sout	sout	thought	N	N	Y	N
46	The new philm is boring for most audience	philm	film	Y	Y	Y	Y
47	It seems as zow most people just ignore this problem	zow	though	N	N	N	N
48	She was accused for asalting her own employee	asalting	assaulting	Y	Y	Y	Y
49	The tension in the team is not a seacreate anymore	seacreate	secret	Y	Y	N	N
50	The existance of the sron is a mystery for centuries	sron	throne	N	N	N	N
51	Not ivorywon can accomplish such a goal	ivorywon	everyone	N	Y	Y	Y
52	The sun sets over the calm, blew ocean	blew	blue	Y	Y	Y	Y
53	A lone wolf houwles beneath the full moon	houwles	howls	Y	Y	Y	Y
54	The closet is not a plays for lost toys	plays	place	Y	Y	Y	Y
55	Soft rain taps ahgaynst the windowpane	ahgaynst	against	Y	Y	Y	Y
56	Books transpurte us to different worlds	transpurte	transport	Y	Y	Y	Y
57	Stars twinkle in the mihdknyte sky above	mihdknyte	midnight	N	Y	N	Y
58	Coffee warms my hahndes on a cold morning	hahndes	hands	Y	Y	Y	Y
59	Time flys when you're having fun	flys	flies	Y	Y	Y	Y
60	Birds cherpe happily in the early morning	cherpe	chirp	Y	Y	Y	Y
61	The old oak tree provides shade in someher	someher	summer	N	N	N	N
62	Love blooms like a dellakit flower	dellakit	delicate	N	N	N	Y
63	Dreams carry us beeyohnde reality's grasp	beeyohnde	beyond	Y	Y	Y	Y
64	Music suthes the weary soul's ache	suthes	soothes	Y	Y	Y	Y
65	Snow blaynkitse the town in a soft hush	blaynkitse	blankets	Y	Y	Y	Y
66	Hisstury whispers through ancient ruins	histtury	history	Y	Y	Y	Y
67	Cullers dance across the painter's canvas	cullers	colors	N	Y	N	Y
68	Whisdumb comes with age and experience	whisdumb	wisdom	Y	Y	Y	Y
69	A cozy fyr crackles in the hearth	fyr	fire	Y	Y	Y	Y
70	Misste settles over the tranquil valley	misste	mist	N	Y	Y	Y
71	Ahdvenchur awaits beyond the horizon	ahdvenchur	adventure	Y	Y	Y	Y
72	Joewhee shines in a child's innocent eyes	joewhee	joy	N	Y	N	Y
73	Wayvs crash against the rugged cliffs	wayvs	waves	N	Y	Y	Y
74	Stahres align for those who persevere	stahres	stars	Y	Y	Y	Y
75	Lafftur echoes through the bustling market	lafftur	laughter	N	Y	N	Y
76	Shahdohs lengthen as the day draws to a close	shahdohs	shadows	N	Y	N	Y
77	Cinednis costs nothing but means everything	cinednis	kindness	N	Y	Y	N
78	An open harte embraces life's surprises	harte	heart	Y	Y	Y	Y
79	Knyte descends like a velvet curtain	knyte	night	N	Y	N	N
80	Spreeng blossoms bring hope anew	spreeng	spring	Y	Y	Y	Y
81	Wispurrs carry secrets on the breeze	wispurrs	whispers	Y	Y	N	Y

Figure 5: English Testset 02

80	Spreeng blossoms bring hope anew	spreeng	spring	Y	Y	Y	Y
81	Wispurrs carry secrets on the breeze	wispurrs	whispers	Y	Y	N	Y
82	Peyciance is a virtue worth cultivating	peyciance	patience	Y	Y	Y	N
83	The younivurs reveals its mysteries slowly	younivurs	universe	Y	Y	Y	Y
84	Twomarrow holds endless possibilities	twomarrow	tomorrow	Y	Y	Y	Y
85	Wimsee guides us to unexpected places	wimsee	whimsy	N	Y	N	N
86	Reefleckcians shimmer on the tranquil lake	reefleckcians	reflections	N	Y	N	Y
87	In sylanse, the mind finds clarity	sylanse	silence	N	Y	N	N
88	Kurradge is found within, not without	kurradge	courage	N	Y	N	Y
89	Nothing sounds better than a nice cup of coughey	coughey	coffee	Y	Y	Y	Y
90	Everything is falling ahpheart and there is nothing we can do	ahpheart	apart	Y	Y	Y	Y
91	Harrmoney weaves through melodies and hearts	harrmoney	harmony	Y	Y	Y	Y
92	Ays mirror the depths of the soul	ays	eyes	Y	Y	N	Y
93	Thunndar rumbles in the distant storm	thunndar	thunder	Y	Y	Y	Y
94	Thyme heals wounds and writes stories	thyme	time	Y	Y	N	Y
95	Summer reign refreshes the thirsty earth	reign	rain	Y	Y	N	Y
96	The mihstaree will not be solved by standing around	mihstaree	mystery	N	Y	N	Y
97	In youknity, strength emerges effortlessly	youknity	unity	Y	Y	N	Y
98	Mines connect across distances through words	mines	minds	Y	Y	N	N
99	Awetumm leaves dance in the crisp breeze	awetumm	autumn	N	Y	N	Y
100	Saraknity comes from within, not without	saraknity	serenity	N	N	Y	Y
101	Each day is a new chahptar waiting to unfold	chahptar	chapter	N	Y	Y	Y

Figure 6: English Testset 03

1	sentence	word	answer_english	answer_spanish	GPT-3.5	GPT-4	Ernie-3.5	Ernie-4.0
2	La muller trabaja de su casa	muller	woman	mujer	Y	Y	Y	Y
3	Quiero ir ha todos los festivales estacionales con mis amigos este verano	ha	to	a	Y	Y	N	N
4	Asia del sur, habrá una tormenta grande en dos días	Asia	toward	Hacia	N	N	N	Y
5	Cuando el habrió la puerta, vi un elefante grande	habrio	she/he opened	abrio	Y	Y	N	Y
6	Su haya las lented que se pierde	haya	find	halla	N	N	N	N
7	Su creio gravar la película por luego	gravar	record	grabar	Y	Y	Y	N
8	Ella hora a su dio por llueve	hora	pray	ora	Y	Y	Y	Y
9	La mujer vieja savia vivía en un bosque sobre una montaña grandiosa	savia	she/he knew	sabia	N	Y	N	N
10	Vaya! Sus votas son que lindisimo, quiero llorar	votas	boots	botas	N	Y	N	N
11	Si viaja a Francia y paga los gentes correcto, puede recibir un jira extraño	jira	tour	gira	N	Y	N	N
12	El pescador reparó la maya de su red antes de lanzarla al mar para asegurarse de no perder ninguna captura	maya	net	mallá	N	Y	N	N
13	La honda del mar golpeó suavemente la orilla mientras los niños jugaban en la arena	honda	wave	onda	N	Y	N	N
14	El acerbo cultural de la región es impresionante, con una rica historia que se refleja en sus museos, música y festivales	acerbo	heritage	acervo	N	Y	N	N
15	As hecho un excelente trabajo con el proyecto de ciencias	as	you have	has	Y	Y	N	Y
16	El basto océano esconde misterios aún no descubiertos por el ser humano	basto	vast	vasto	Y	Y	N	N
17	Después de la pubertad, el bello corporal comenzó a aparecer en diferentes partes de su cuerpo, marcando una etapa de cambios físicos evidentes	bello	hair	vello	N	Y	N	N
18	Si bienes a la fiesta, asegúrate de traer tu juego de mesa favorito	bienes	you come	vienes	Y	Y	N	N
19	La versión veta del software estará disponible para los usuarios que se registren para las pruebas antes del lanzamiento oficial	veta	beta	beta	Y	Y	N	N
20	Cuando intentaba alcanzar la fruta más alta, Juan cayó del árbol y se lastimó el tobillo	caió	she/he fell	cayó	Y	Y	N	N
21	Es humano herrar, pero solo los sabios aprenden de sus errores	herar	to err	errar	Y	Y	N	Y
22	La situación económica del país es grave y requiere medidas urgentes.	grave	grave	grave	Y	Y	N	N
23	Después de a ver revisado todos los documentos, encontramos el error que causaba el problema	a ver	to have	haber	Y	Y	N	N
24	Durante la ceremonia, el soldado llevó con orgullo el hasta de la bandera al frente del desfile	hasta	pole	asta	Y	Y	N	N
25	Antes de salir de la librería, decidí hojear rápidamente algunos títulos en la sección de novedades	hojear	to browse	ojear	Y	Y	N	N
26	Cuando llegamos a la playa, la primera hola nos mojó los pies mientras caminábamos por la orilla	hola	wave	ola	Y	Y	N	N
27	Yo huso una computadora nueva para mis clases en línea	huso	I use	uso	Y	Y	N	N
28	Cuando llegué a casa, mi madre ya había puesto la sopa a cocer en la hoya	hoya	pot	olla	Y	Y	N	N
29	Para preparar la ensalada, primero necesitas rayar la zanahoria y el queso	rayar	to grate	rallar	Y	Y	N	N
30	El científico decidió rebelar sus hallazgos en la conferencia internacional para compartir su descubrimiento con el resto de la comunidad académica	rebelar	to reveal	revelar	Y	Y	N	N
31	Durante nuestras vacaciones, disfrutamos de un tranquilo paseo por la ribera del río, observando cómo el agua reflejaba el brillante sol de la tarde	ribera	river bank	rivera	Y	Y	N	N
32	Es fundamental aprender los conceptos básicos de la programación para poder avanzar en estudios más complejos de informática	aprender	to apprehend	aprehender	N	N	N	N
33	En el jardín de mi abuela crece hacadera, la cual utiliza para preparar una deliciosa sopa que sirve como remedio casero para la salud	hacadera	dock	acadera	N	N	N	N
34	La baca en el campo estaba muy tranquila pastando bajo el sol de la tarde	baca	cow	vaca	Y	Y	N	N
35	El paisaje montañoso era tan vello que todos los turistas se detuvieron a tomar fotos	vello	beautiful	bello	Y	Y	N	N
36	Antes de salir de compras, revisa cuánto bale ese producto en diferentes tiendas para asegurarte de obtener el mejor precio	bale	okay	vale	Y	Y	N	N
37	El campesino utilizó un vasto para montar su caballo durante el trabajo en el campo	vasto	coarse	basto	N	Y	N	N
38	Después de la tormenta, el jardín quedó completamente desecho, con flores y ramas esparcidas por todas partes	desecho	undone	deshecho	N	Y	N	N
39	Después de mucho esfuerzo, he echo mi tarea de matemáticas completamente solo	echo	she/he made	hecho	Y	Y	N	N
40	La precedencia de los ingredientes es crucial para garantizar la calidad del plato en un restaurante gourmet	precedencia	origin	procedencia	N	N	N	N
41	Durante el otoño, los niños se divierten al deshojar las margaritas para adivinar si les corresponde el amor de su amado	deshojar		desojar	N	N	N	N
42	Durante nuestra caminata por el bosque, nos detuvimos a admirar un enebro centenario cuya majestuosidad destacaba entre los demás árboles	enebro	juniper	enebro	Y	Y	N	N

Figure 7: Spanish Testset 01

40	La precedencia de los ingredientes es crucial para garantizar la calidad del plato en un restaurante gourmet	precedencia	origin	procedencia	N	N	N	N
41	Durante el otoño, los niños se divierten al deshojar las margaritas para adivinar si les corresponde el amor de su amado	deshojar		desojar	N	N	N	N
42	Durante nuestra caminata por el bosque, nos detuvimos a admirar un enebro centenario cuya majestuosidad destacaba entre los demás árboles	enebro	juniper	enebro	Y	Y	N	N
43	La encima lactasa ayuda a descomponer la lactosa, lo que permite a las personas con intolerancia a la lactosa digerir productos lácteos más fácilmente	encima	ensyme	enzima	Y	Y	N	N
44	La coneja en el jardín estaba buscando zanahorias para alimentar a sus pequeños	consejo	doe	coneja	N	Y	N	N
45	Mi postre favorito es una tarta de presas fresca y deliciosa	presas	strawberry	fresa	Y	Y	N	Y
46	El jardinero usó una pala para cavar el hoyo donde plantaría el nuevo árbol en el jardín	pala	shovel	pala	N	Y	N	N
47	Después de comer, Juan lavó los platos usando una esponja nueva para asegurarse de que quedaran completamente limpios	econja	sponge	esponja	N	Y	N	N
48	Después de la caída, Juan sintió un dolor intenso en el brazo y decidió visitar al médico para asegurarse de que no estaba fracturado	abrazo	arm	brazo	N	Y	N	N
49	En nuestras vacaciones de verano, tuvimos la oportunidad de ver un pulpa en su hábitat natural mientras buceábamos cerca de la costa	pulpa	octopus	pulpo	Y	Y	N	N
50	Los científicos continúan explorando la posibilidad de vida en otros planetas del sistema solar	plantas	planets	planetas	Y	Y	N	N
51	Recibí una carga de mi mejor amigo que se mudó al extranjero el año pasado	carga	latter	carta	N	N	N	N
52	La luna plena iluminaba el camino a través del bosque, creando un paisaje mágico y sereno	plena	full	llena	Y	Y	N	N
53	En el zoológico, la serpiente pitón atrajo a muchos curiosos con su impresionante longitud y sus movimientos sinuosos	serpentina	snake	serpiente	Y	Y	N	N
54	Por causalidad, encontré un viejo amigo en la librería mientras buscaba un libro que hacía años quería leer	causalidad	chance	casualidad	Y	Y	N	N
55	El profesor siempre mostraba paciencia al explicar los conceptos más complejos a los estudiantes más inductos	inductos	unlearned	indoctos	N	N	N	N
56	Decidimos alquilar un apartamento cerca de la playa para disfrutar del verano	alquilar	to rent	alquilar	Y	Y	Y	Y
57	Es importante no divergir las palabras de los demás para mantener una comunicación clara y honesta	divergir	to distort	tergiversar	Y	Y	N	N
58	El perito de su éxito se debe a su dedicación y esfuerzo constante en el proyecto	perito	merit	mérito	N	N	N	N
59	Es importante cembra árboles para ayudar a mantener el equilibrio ecológico y mejorar la calidad del aire que respiramos	cembra	to sow	sembrar	Y	Y	N	N
60	El mecánico diagnosticó que la falla del motor se debía a un problema con la bomba de combustible	falla	failure	falla	Y	Y	N	N
61	Después de muchos años de servicio, el profesor anunció su retiro del ámbito académico para disfrutar de su jubilación	retirado	she/he withdraws	retirada	Y	N	N	N
62	La chef del departamento anunció hoy los nuevos planes de expansión de la empresa durante la reunión	chefa	boss	jefa	Y	Y	N	N
63	El mecánico tuvo que arranque el motor varias veces antes de que finalmente comenzara a funcionar correctamente	arranque	to start up	arrancar	Y	Y	N	N
64	El agua es un concurso natural esencial para la vida en nuestro planeta	concurso	resource	recurso	N	Y	N	N
65	Durante el concierto, la plateada estaba completamente llena, y el entusiasmo del público se sentía en el aire	plateada	orchestra	platea	Y	Y	N	N
66	Costa Rica es conocida por su pica biodiversidad y hermosos paisajes naturales	pica	delicious	rica	Y	Y	N	N
67	Durante el festival, instalaron una enorme arpa en el parque para albergar a los músicos y a los asistentes	arpa	large tent	carpa	Y	Y	N	N
68	La demostrador en la tienda fue muy eficaz al explicar cómo funcionaban los nuevos electrodomésticos	demonstrador	demonstrator	demonstradora	Y	Y	N	N
69	Voy a arengar el reloj que se descompuso la semana pasada	arengar	to arrange	arreglar	Y	Y	Y	Y
70	En la clase de español, el profesor pidió a los estudiantes que practicaran cómo deletrear correctamente las palabras difíciles	deleitar	to spell	deletrear	N	Y	N	N
71	Senti una gran berenjena cuando me di cuenta de que había olvidado el nombre de mi compañero de trabajo durante la presentación	berenjena	shame	vergüenza	N	Y	N	N
72	El casto en tecnología este año ha superado el presupuesto inicial debido a la necesidad de actualizar nuestros sistemas informáticos	casto	spent	gasto	Y	N	N	N
73	Durante la feria internacional, el campeón de tecnología fue el más visitado, atrayendo a entusiastas de todo el mundo	campeón	pavilion	pabellón	N	N	N	N
74	Durante nuestra visita al museo, el ría nos explicó detalladamente la historia de cada obra de arte	ría	guide	guía	Y	Y	N	N
75	El pastel de chocolate que preparó mi abuela era increíblemente dulce y delicioso	duce	sweet	dulce	Y	Y	N	N
76	La carretera mojada se volvió pilosa después de la tormenta	pilosa	dangerous	peligrosa	N	Y	N	N
77	Cuando abrí la cortita esta mañana, la luz del sol inundó la habitación, llenándola de calidez y energía	cortita	curtain	cortina	N	Y	N	N
78	Su aditamento al hablar frente a una audiencia tan grande fue admirado por todos los presentes	aditamento	dare	atrevimiento	N	N	N	N
79	Mi amigo me va a surgir algunos libros interesantes para leer durante las vacaciones	surgir	to suggest	sugerir	Y	Y	N	N
80	Es importante ser consistente de nuestras acciones y cómo estas afectan a los demás	consistente	aware	consciente	N	N	N	N
81	El agricultor compra adorno orgánico cada primavera para mejorar la fertilidad de su tierra y obtener una mejor cosecha	adorno	she/he passes	abono	Y	Y	N	N

Figure 8: Spanish Testset 02

80	Es importante ser consistente de nuestras acciones y cómo estas afectan a los demás	consistente	aware	consciente	N	N	N	N
81	El agricultor compra adorno orgánico cada primavera para mejorar la fertilidad de su tierra y obtener una mejor cosecha	adorno	she/he passes	abono	Y	Y	N	N
82	El guerrero varvaro avanzaba con fiereza a través del campo de batalla, dejando a su paso un rastro de valentía y temor	varvaro	barbarian	bárbaro	Y	Y	N	N
83	Desde la cima de la montaña, la pista era absolutamente impresionante, abarcando valles y ríos que brillaban bajo el sol del atardecer	pista	view	vista	Y	N	N	N
84	Los niños construyeron un castillo de arenga en la playa durante sus vacaciones de verano	arenga	sand	arena	Y	Y	N	N
85	Antes de visitar la exposición de arte, decidí leer el boletín que ofrecían en la entrada para entender mejor las obras y artistas presentados	boletín	information	folleto	Y	Y	N	N
86	Mi hermano juega al balonmano todos los sábados en el parque local	balonmano	basketball	baloncesto	N	N	N	N
87	La columna entregó su proyecto final de ciencias con gran entusiasmo y recibió elogios de su profesor	columna	alumni	alumna	N	Y	N	N
88	El programa nacional de educación busca erradicar el analfabetismo y mejorar el alfabetización funcional entre los adultos de la región	alfabetización	literacy	alfabetismo	Y	Y	N	N
89	El collar que llevaba puesto tenía incrustado un diametral que brillaba intensamente bajo la luz del sol	diametral	diamond	diamante	Y	Y	N	N
90	El juez dio la horden de detener al sospechoso después de revisar todas las pruebas presentadas en el caso	horden	they order	orden	Y	Y	N	N
91	Para conservar mejor las verduras, es recomendable cohelarlas inmediatamente después de comprarlas	cohelarlas	to freeze them	congelarlas	Y	Y	N	N
92	La bicicleta no puede avanzar porque tiene la rueda pinchada	hueda	wheel	rueda	Y	Y	Y	Y
93	El principal espectáculo para alcanzar nuestras metas es a menudo nuestro propio miedo al fracaso	espectáculo	obstacle	obstáculo	N	Y	N	N
94	Durante el invierno, la refacción de nuestra casa es esencial para mantenernos cálidos y confortables	refacción	heating	calefacción	N	N	N	N
95	El argumento en las ventas este trimestre ha superado todas nuestras expectativas	argumento	increase	aumento	N	Y	N	N
96	La dispersar de semillas por los animales es crucial para la regeneración de muchos ecosistemas forestales	dispersar	dispersion	dispersión	N	Y	N	N
97	Guardamos todas las herramientas de jardinería en el obertizo para protegerlas de la lluvia y el sol	obertizo	shed	cobertizo	Y	Y	N	N
98	El primer semestre del año escolar siempre es crucial para establecer un buen ritmo de estudio	semestre	quarter	trimestre	N	N	N	N
99	Después de la caída, el jarrón quedó completamente quedado en el suelo	quedado	broken	quebrado	Y	Y	N	N
100	Durante la noche oscura, apenas se podían ver las estrellas debido a las nubes que cubrían el cielo	oscura	dark	oscura	Y	Y	N	N
101	El cazador se prepara para la temporada de casa de ciervos	casa	hunt	caza	Y	Y	N	N

Figure 9: Spanish Testset 03