



EXPLORING CHATGPT'S ACCURACY AND CONFIDENCE IN HIGH-RESOURCE LANGUAGES

Bachelor's Project Thesis

Martino Pelucchi, s4378857, m.pelucchi@student.rug.nl,

Supervisor: Dr M.A. Valdenegro Toro

Abstract: As soon as it was published, ChatGPT took the world by storm for its impressive abilities. Due to its release without documentation, scientists immediately began to attempt to formally identify its limits, mainly through its performance in natural language processing (NLP) tasks. This paper aims to join the growing literature regarding ChatGPT's abilities by focusing on its performance in high-resource languages as well as on its capacity to predict its answers' accuracy by giving a confidence level. The analysis of high-resource languages is of interest as studies have shown that low-resource languages perform worse than English in NLP tasks, but no study so far has analysed whether high-resource languages perform as well as English. The analysis of ChatGPT's confidence calibration has not been carried out before either and is critical to learn about ChatGPT's trustworthiness. In order to study these two aspects, five high-resource languages and two NLP tasks were chosen. ChatGPT was asked to perform both tasks in the five languages and to give a numerical confidence value for each answer. The results show that all the selected high-resource languages perform similarly and that ChatGPT does not have a good confidence calibration, often being overconfident and never giving low confidence values.

1 Introduction

1.1 ChatGPT

ChatGPT is a large language model developed by OpenAI * based on their previous model GPT-3 †. As a generative pre-trained model, GPT-3 was created by training an extremely large neural network (175 billion parameters) through deep learning on enormous amounts of data (Brown et al., 2020). Thanks to its generative nature it is able to create coherent text starting from an initial input. ChatGPT is a fine-tuned version of this model specialising in following instructions and holding conversations. Moreover, it was also trained using reinforcement learning from human feedback (Christiano et al., 2017), which leads to more accurate and faster learning thanks to improved reward functions.

Ever since its release on the 30th of November 2022, ChatGPT has become extremely popular due to its accessibility and ability to provide answers to all sorts of requests. In the first week, the chatbot developed by OpenAI reached a million users and as of January 2023, it surpassed 100

million users (Ruby, 2023). It is until now freely available and has a very straightforward user interface. As it gained popularity, users quickly realised its impressive natural language processing (NLP) and multilingual abilities, but as the model was released without documentation, its limits were not clear. To better understand ChatGPT's strengths and weaknesses, researchers started to analyse the model's performance scientifically.

1.2 Related Work

ChatGPT was compared to a multitude of models, from fine-tuned to general ones, in a variety of NLP tasks and languages.

Qin et al. (2023) used 7 different tasks (reasoning, natural language inference, question answering, dialogue, summarization, named entity recognition, and sentiment analysis) to compare ChatGPT to GPT 3.5 and to models fine-tuned to the tasks. They found that ChatGPT performed slightly worse than fine-tuned models in the majority of tasks, but its performance was better than GPT 3.5 for the majority of tasks except for reasoning tasks like common sense questions.

*<https://openai.com/>

†Updated to GPT 3.5

Jiao et al. (2023) analysed its multilingual abilities through translation tasks between English, German, Chinese, and Romanian. The results indicated that ChatGPT performs similarly to popular machine translation models like Google Translate for high-resource languages such as English and German, but falls behind when translating into Romanian, a lower-resource language.

In order to study ChatGPT’s NLP abilities and language knowledge Bang et al. (2023) performed a multilingual and multitask evaluation. They performed 7 NLP tasks (summarisation, machine translation, sentiment analysis, question answering, misinformation detection, task-oriented dialogue, open domain kgd) across a set of languages with mixed resource levels (English, Chinese, French, Indonesian, Korean, Javanese, Sundanese, Buginese). Although the results slightly varied depending on the task, the general finding was that ChatGPT achieved strong performances in high- and medium-resource languages (e.g. English, Indonesian) while it struggled to understand low-resource languages (e.g. Javanese, Sundanese).

These papers indicate that although ChatGPT can understand a vast number of languages, its performance on NLP tasks is dependent on the amount of resources it has at its disposal for a specific language. In order to study ChatGPT’s performance when executing tasks in different languages, it is therefore necessary to look at their resource levels. To do so, the spread of languages in ChatGPT’s training data must be investigated.

1.3 Linguistic Resources

OpenAI keeps the data that was used to train any of their GPT models confidential; it is therefore impossible to know exactly the percentage of texts in each language in the training data. Nonetheless, there are a few heuristics that can be used to get an educated guess on the matter. Although the specifics are not public, it is generally thought that OpenAI gets its datasets from crawling the internet, but not much information is known. Consequently, statistics such as the distribution of languages in web content (Petrosyan, 2023) or in the Common Crawl dataset (commoncrawl, n.d.) can give an estimation of the distribution of language in ChatGPT’s training data.

Table 1.1 gives an overview of the percentage of

Table 1.1: Distribution of the most popular European languages in web content (Petrosyan, 2023) and in the Common Crawl dataset (commoncrawl, n.d.) in percentage.

Languages	Web Content	Common Crawl
English	58.8%	46.3%
Spanish	4.3%	4.6%
German	3.7%	5.9%
French	3.7%	4.7%
Italian	1.6%	2.7%

web content and of the Common Crawl dataset in different European languages. Although, for example, 1.6% might seem like a very small percentage, all the languages listed are still commonly considered high-resource, which is usually defined as languages with over 1% of web content. In fact, ChatGPT seems fluent in all five languages. Nevertheless, the disparity between English and other languages is still notable. This begs the question of whether a more in-depth analysis would show that the disparity in language distribution leads to a significant difference in performance.

1.4 Scope of the Research

Although ChatGPT’s language processing abilities have been studied through NLP tasks, very little research has been carried out on whether non-English high-resource languages show the same level of performance as English. The research on ChatGPT’s multilingual abilities has focused on Machine Translation, which does not lead to any conclusions on its abilities to perform NLP tasks in different languages. Considering ChatGPT was released with little to no documentation, research is the only approach to discovering the limits and strengths of its linguistic abilities. It is therefore necessary to perform thorough testing and analyse ChatGPT’s abilities in different languages.

Another aspect of ChatGPT that has not yet been studied is its confidence, namely whether it is able to give a reliable confidence value for its answers, one that matches their accuracy. The study of ChatGPT’s confidence calibration is important due to the blind trust many users have in its answers. Simple interactions with the chatbot show that it often gives inaccurate answers rather than

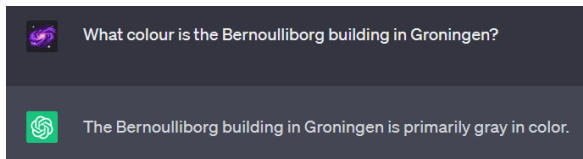


Figure 1.1: ChatGPT answering with an incorrect fact rather than admitting its lack of knowledge on the topic. The building in question is primarily blue.

acknowledging its ignorance on the topic, as shown in Figure 1.1.

However, when prompted to, ChatGPT is able to give a level of confidence in its answer, but no study has analysed whether its perception is accurate. Discovering whether ChatGPT gives a reliable confidence value on its answers would be a step forward towards being able to recognise which answers can be trusted.

This paper aims to fill the gap in the literature by comparing ChatGPT’s accuracy and confidence across multiple high-resource languages in different NLP tasks. In doing so, it strives to answer two research questions:

- Is ChatGPT equally accurate in high-resource languages?
- Is ChatGPT’s confidence in its answers well-calibrated?

In the following sections the tasks, languages and data will be discussed, as well as the setup of the experiment. Afterwards, the results will be presented and discussed. Lastly, conclusions will be drawn.

2 Methods

2.1 Tasks and Languages

In order to study ChatGPT’s accuracy and confidence in high-resource languages, it was decided to compare its performance in two NLP tasks and five languages. The tasks were sentiment analysis (SA) and common sense reasoning (CSR), while the languages were English, Spanish, German, French and Italian.

Sentiment analysis is a task whose goal is to find opinions and identify the sentiment they carry, whether positive, negative, or neutral (Medhat et

al., 2014). Common sense reasoning is the task of making inferences or answering questions based on common sense knowledge. This knowledge is considered to be the set of facts that reflect the general understanding of the world and of humans (Davis & Marcus, 2015). These tasks were chosen as they involve two skills that are fundamental for conversation and consequently for chatbots like ChatGPT.

Previous literature was analysed to decide which languages to use in this study. Since the only paper which performed a multilingual and multitask analysis of ChatGPT focused on Asian languages (Bang et al., 2023), European languages were chosen as the focus of this study. The decision to use the most popular languages was based on multiple factors. The first one is the number of speakers. Using the most popular languages means that the findings of this paper can affect a high number of people and their usage of ChatGPT. The second factor is that there is more uncertainty around non-English high-resource languages. While the drop in performance for low-resource languages is obvious, it is not clear whether ChatGPT’s performance in other high-resource languages is as reliable as in English. In fact, although all five languages are considered high-resource, there is still a notable difference in resources between the languages.

2.2 Data

To study ChatGPT’s performance in different languages it was necessary to find multilingual datasets for the two tasks of sentiment analysis and common sense reasoning. To this end, Hugging Face was used. Hugging Face[‡] is an AI Community that promotes open-source contributions and acts as a hub for models and datasets. The datasets can be filtered based on tasks and languages and this option was used to select the two datasets for this project.

In order to study how ChatGPT performs in SA, the Unified Multilingual Sentiment Analysis Benchmark (UMSAB) (Barbieri et al., 2022) was chosen. This multilingual dataset was constructed by combining 8 monolingual sentiment analysis datasets. All datasets are made of 3033 tweets evenly split between the three possible labels: "positive", "negative" and "neutral".

[‡]<https://huggingface.co/>

For CSR, the XCSR dataset (Lin et al., 2021) was chosen. This dataset was created to improve multilingual language models in CSR. It consists of common sense multiple-choice questions with 5 possible answers in 16 languages. The dataset is split into two sub-datasets: X-CODAH and X-CSQA. The validation split of X-CSQA was chosen, as it is the only part which contains all the necessary fields. For each language, it contains 1000 datapoints.

2.2.1 Preprocessing of data

Although the two original datasets are already curated, some preprocessing steps were needed for the scope of this project.

Firstly, as this experiment did not envisage a validation phase, the data splits were merged together, in order to have a larger pool of sentences and questions.

Secondly, in the UMSAB dataset, some sentences contained Unicode characters that had not been converted. All the sentences containing a Unicode character were excluded to remove any possible confusion for ChatGPT.

Thirdly, some sentences and questions around topics such as sex and drugs caused ChatGPT to give warnings regarding its content policy. It was not possible to know which sentences would violate the content policy and to remove them beforehand automatically. Therefore, those sentences or questions were removed manually as they were found.

Lastly, as the datapoints needed to be manually fed to ChatGPT in small batches, the size of the datasets was reduced. For every language and task, a subset of 300 datapoints was randomly sampled. For the sentiment analysis task, the subset was kept balanced.

2.3 Prompt Engineering

As a chatbot, the only way to interact with ChatGPT is through prompts. Therefore, a prompt needed to be established for each task. The prompt had to be defined carefully since it had multiple roles. Firstly, the task needed to be explained. For this experiment, this meant specifying how to classify the sentiment of tweets or how to answer multiple-choice questions. Secondly, the format of the input needed to be explained. This was done to ensure that ChatGPT knew what to expect and

what to do with the tweets or questions. Lastly, the output desired from ChatGPT needs to be defined. This was done to allow for an automatic scanning and analysis of the responses.

The process of creating the prompt started with simple requests to perform sentiment analysis or answer multiple-choice questions. ChatGPT immediately showed an understanding of the tasks, but the format of its response varied. The answer was often explained and the confidence was not expressed numerically. This led to the addition of a sentence detailing the desired output. After noticing that an in-depth explanation of the task did not seem to influence the answers, the focus shifted to the number of datapoints that could be sent together (a block).

The goal was to maximise the datapoints per block to speed up the experiment process. At around 20 tweets per block ChatGPT would often “forget” the task and simply describe the tweets (e.g. saying that they are a collection of tweets). To avoid this issues, a sentence describing the block that was to come was added and its size was restricted to 15 datapoints.

Lastly, the English prompt was translated by a professional translator into the different languages in an attempt to improve the performance in non-English languages. A small experiment was conducted using sentiment analysis to compare the performance of ChatGPT using the English versus the translated prompts. Since a significant difference in accuracy was found for some languages (see Appendix A.1), the translated prompt was used for all languages and tasks.

After testing out different variations, the final English prompts that were used for the experiment were the following:

- **Sentiment Analysis:** Hello, I would like you to perform sentiment analysis on 15 English sentences. Please classify each sentence as “Positive”, “Negative”, or “Neutral” according to the sentiment it expresses. Moreover, please rate your confidence in the answer you gave between 0 and 100%. The answer should be a list with the format “Sentiment (confidence%)”. Say “understood” if you have understood. The list of sentences will follow.
- **Common Sense Reasoning:** Hello, I would like you to answer 15 common sense multiple-

choice questions in English. For each question please give the letter that corresponds to the correct answer. Moreover, please give your confidence in the answer you gave between 0 and 100%. The answer should be a list with the format “Letter (confidence%)”. Say “understood” if you understand. The list of questions will follow.

A complete list of all the prompts used in the different languages for both tasks can be found in Appendix A.2.

2.4 Exploratory Analysis

Once the prompt was defined, an exploratory analysis was conducted. Its goal was to identify patterns in the responses as well as any confounding variables.

A first finding was that, after the prompt, the number of blocks that could be sent consecutively was not consistent. Sometimes ChatGPT would perform the task on 3 blocks, while in other cases it stopped after the first block. It was therefore decided to send the prompt before each block.

A second finding was discovered when sending the same block multiple times. It was found that ChatGPT’s answers varied, not only in confidence but also in the sentiment or answer chosen. Due to this discovery, it was decided to repeat the experiment 3 more times with a subset of the data in order to analyse the fluctuation in accuracy and confidence.

Another factor that was taken into account and tested is that ChatGPT’s memory is chat-specific. The tests found that the same datapoints would sometimes lead to different answers even in the same chat. Nonetheless, in order to avoid any possible influence of the memory, a new chat was created for each language and for the repetition step of the experiment.

2.5 Procedure

The experiment was run on OpenAI’s website on the May 24 version of ChatGPT. The use of the website rather than the API was initially due to the non-existence of the API, and it was later preferred as most people use the web version of ChatGPT.

The process of gathering the answers was a repetition of the same two steps for each language and

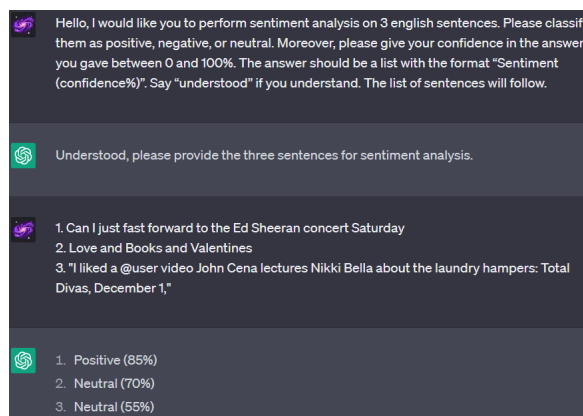


Figure 2.1: Showcase of the experiment procedure with ChatGPT. Number of datapoints lowered to 3 for demonstration purposes.

task. The prompt was sent and then a block of 15 sentences or questions followed. As there were 300 datapoints, the process was repeated 20 times per task and per language, every time in a new chat. The list of answers for each block was then copied into a text file in order to assemble the answers for all the blocks. An example of the “conversation” can be seen in Figure 2.1, while more examples from all tasks and languages can be found in Appendix A.3.

Afterwards, for every language and task, a subset of 75 datapoints was created. These subsets were sent 3 times each. The sentences and questions were presented in the same way to ChatGPT as previously explained.

3 Results

3.1 Sentiment Analysis

3.1.1 Qualitative Results

In order to plot the results of the SA task, the data was grouped in 20 bins, each representing a 5% confidence range. In Figure 3.1, a calibration plot showcasing the accuracy per confidence bin of each language in the sentiment analysis task is presented. Before analysing it, the density of each confidence bin must be considered (see Figure 3.2). The first important result from the graphs is that ChatGPT does not give low confidence values, with the lowest being 50% (60% for Spanish) and the mode

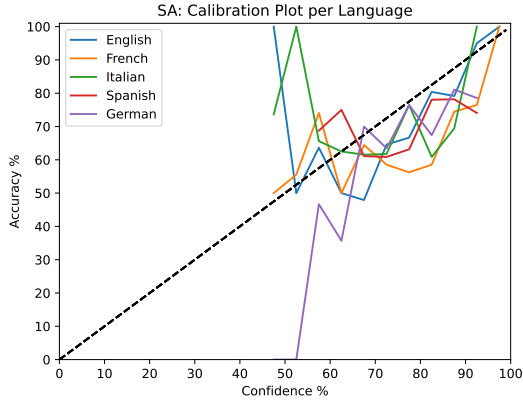


Figure 3.1: Calibration plot of each language for the sentiment analysis task with the diagonal line showing the perfect calibration. ChatGPT is mostly overconfident in all languages.

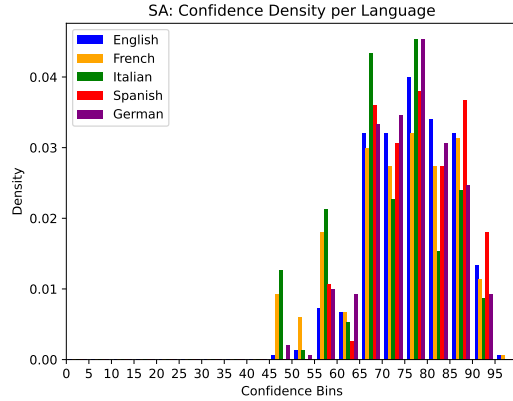


Figure 3.2: Density histogram of each language’s confidence bins for the sentiment analysis task. ChatGPT does not give a confidence level below 50% in any language.

being 80% across all languages. When focusing on the calibration plot, no language seems to perform better than the others, and it appears that all languages approximately follow the diagonal line representing a perfect calibration. Nonetheless, when declaring a confidence value between 50 and 60%, ChatGPT appears underconfident (except for German), whereas between 60 and 90% it appears overconfident. In the languages (English, French and Italian) that contain answers with confidence above 90%, ChatGPT is appropriately confident. Pairing these results with the density of the bins shows that the high accuracies for low confidences may be due to the low density in the concerned bins and that where most answers lie ChatGPT is overconfident.

Furthermore, an accuracy-confidence curve was generated to further analyse ChatGPT’s reliability, shown in Figure 3.3. The curve was created by removing the data below an increasing threshold and computing the mean accuracy of the remaining data. It shows that, for all languages, the accuracy does not increase when the threshold increases from 50 to 70%. In fact, the accuracy remains at around 65% for French and 70% for all other languages. This could be due to the general decrease in accuracy in that confidence range, as shown in Figure 3.1. After the 70% confidence threshold, all languages show some improvements as the threshold increases, with English and French having the biggest and smallest improvement re-

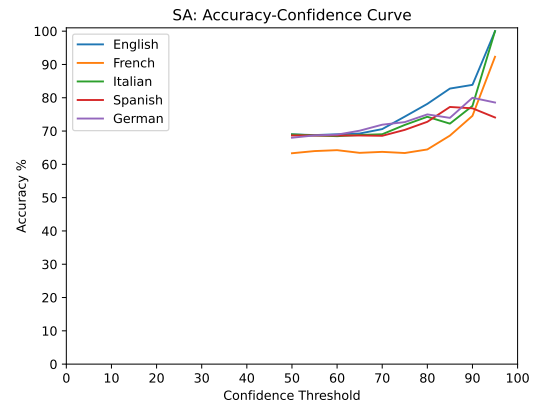


Figure 3.3: Accuracy-confidence curve of every language for the sentiment analysis task. Increasing the threshold of the data considered has very little influence on the accuracy.

spectively. Lastly, when the thresholds are set to 90 and 95%, English, Italian and French show a steep increase in accuracy, while German and Spanish a slight decrease.

3.1.2 Quantitative Results

The accuracy and confidence of ChatGPT in the SA task in the different languages can be found in Table 3.1. English and Italian have the highest accuracy at 69%, with Italian also having the lowest

Table 3.1: Summary of the results per language of ChatGPT in the sentiment analysis task. The values shown are mean accuracy and confidence (with standard deviation) as well as expected calibration error and maximum calibration error. There is no significant difference in accuracy between the languages.

Language	Accuracy \pm SD	Confidence \pm SD	ECE	MCE
English	69.0 \pm 1.5	79.3 \pm 2.1	11.1	50.0
French	63.3 \pm 2.4	76.4 \pm 1.8	15.6	26.2
Italian	69.0 \pm 2.0	74.9 \pm 3.3	11.1	45.0
Spanish	68.7 \pm 2.7	80.0 \pm 1.6	12.6	21.0
German	68.0 \pm 2.0	78.1 \pm 1.8	10.1	55.0

confidence. A chi-squared test of independence was performed in order to study the relationship between the language and the accuracy of ChatGPT. The test showed no significant association between the two variables, $\chi^2(4, N = 1500) = 3.208, p = 0.52$.

To get a better understanding of the abilities of ChatGPT, its accuracy in the SA task was compared to that of two other models on the same dataset. The models, XLM-R (Conneau et al., 2020) and XLM-T (Barbieri et al., 2022), are two large multilingual language models, with the second being fine-tuned on Twitter data. To have the best comparison, the F1 score of the two models when trained with all languages’ training data was used. The F1 score was selected as it was the only metric presented for the two other models. It is important to remember that this is only a rough comparison as the ChatGPT results were only on a subset of the data. The F1 Score of the 3 models can be seen in Table 3.2.

From the results, it appears that change in performance partially depends on the language. All 3 models perform approximately the same in Italian

Table 3.2: The F1 score of ChatGPT in the sentiment analysis task compared to that of two benchmark models (XLM-R, XLM-T) on the UMSAB dataset. ChatGPT’s performance is either equal or worse.

Language	ChatGPT	XLM-R	XLM-T
English	68.3	68.5	70.6
French	62.6	70.5	71.2
Italian	68.6	68.6	69.1
Spanish	67.8	66.0	67.9
German	67.9	72.8	77.4

and Spanish but ChatGPT performs worse than the other two in English, French and German. Overall, ChatGPT had a slightly worse performance compared to other benchmark multilingual language models when doing sentiment analysis.

In order to assess the variance in ChatGPT’s answers, the standard deviation of the accuracy and of the confidence across the 4 iterations of 75 datapoints were computed (Table 3.1). The results show that there is slight variation both in the average accuracy and in the average confidence in all languages. Interestingly, English has the lowest accuracy deviation and the highest confidence deviation while for Spanish it is the opposite. It is important to consider that these results may vary if the questions are always asked in the same chat.

Lastly, to analyse the discrepancy between ChatGPT’s confidence in its answers and their accuracy, the expected calibration error (ECE) and the maximum calibration error (MCE) were computed (Guo et al., 2017). The ECE is a weighted average of the absolute difference between the accuracy and the confidence of every bin and can be computed as follows:

$$ECE = \sum_{m=1}^M \frac{|B_m|}{n} \left| \text{acc}(B_m) - \text{conf}(B_m) \right| \cdot 100, \quad (3.1)$$

where M is the number of bins, B_m is the m th bin, n is the number of datapoints and $\text{acc}(B_m)$ and $\text{conf}(B_m)$ are the average accuracy and confidence of the bin. As ECE uses a weighted average, the small range of confidence levels ChatGPT gives does not influence its value. If all bins were considered, the large number of empty bins would make the calibration error notably higher. The MCE is simply the biggest deviation between accuracy and confidence across all bins and can be computed as

follows:

$$\text{MCE} = \max_{m \in \{1, \dots, M\}} |\text{acc}(B_m) - \text{conf}(B_m)| \cdot 100. \quad (3.2)$$

The two results for each language are displayed in Table 3.1. The two metrics range from 0 to 100 and provide a standardised measure of calibration performance, with values closer to zero representing better calibration. The ECE shows similar results across all languages, at around 13, with German having the lowest (10) and French having the highest (16). This means that on average ChatGPT’s confidence levels are 10 to 16% away from the accuracy. This confirms that ChatGPT has an overall decent calibration although not very accurate. The MCE varies more but has high results across all languages, ranging from 26 in French to 55 in German. This means that, although in general ChatGPT has a decent calibration, for some bins the values are completely wrong. This may also be due to the bins with little data having skewed results, as this metric is not influenced by the size of the bin.

3.2 Common Sense Reasoning

3.2.1 Qualitative Results

To analyse the results of ChatGPT in the CSR task, the same steps were taken as for the SA task. Figures 3.4 and 3.5 show similar trends to those of the first task. No confidences below 55% are given and for the vast majority of the answers, between 65 and 90% confidence, ChatGPT is overconfident, reaching accuracies between 50 and 70%. Nonetheless, in that range, the data roughly follows the diagonal of perfect calibration. For confidences below 65% and above 90%, both extremely low accuracy and high accuracy values are seen, but this can be explained by the small size of the bins. In this task as well all languages seem to perform very similarly.

Further analysis performed with the accuracy-confidence curve found in Figure 3.6 shows that increasing the confidence threshold of the data included has no influence on the accuracy up until 75%. In this interval, English has the highest accuracy at 70% and Italian the lowest at 60%. Once the threshold is increased past 75%, the accuracy in French has a noticeable increase while the other

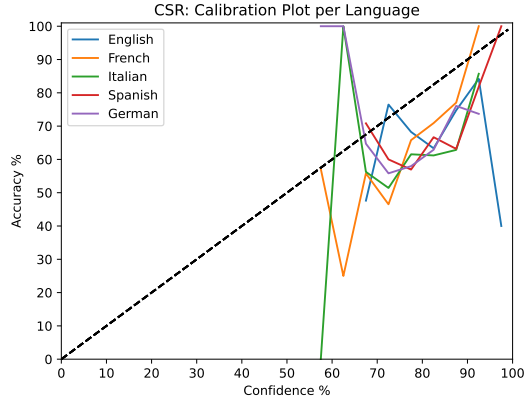


Figure 3.4: Calibration plot of each language for the common sense reasoning task with the diagonal line showing the perfect calibration. ChatGPT is again overconfident in all languages for the majority of the confidence bins.

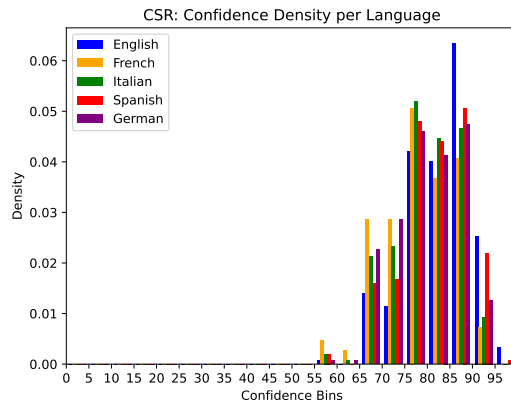


Figure 3.5: Density histogram of every language’s confidence bins for the common sense reasoning task. ChatGPT does not give any confidence levels below 55% in any language.

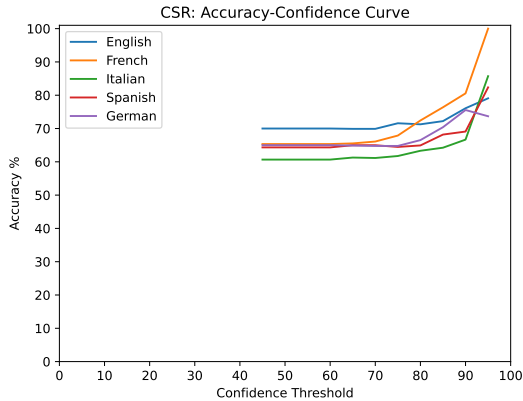


Figure 3.6: Accuracy-confidence curve of every language for the common sense reasoning task. Increasing the threshold of the data considered has no strong effect on the accuracy.

languages only improve slightly until 90%, at which point Spanish and Italian have a steep increase, German has a slight decrease and English continues on the same trend.

3.2.2 Quantitative Results

The accuracy and confidence of ChatGPT in the different languages while performing the CSR task can be found in Table 3.3. English has the highest accuracy (and confidence) while Italian has the lowest. It is also interesting to see that, although the mean confidence is higher than for the SA task, the accuracies are lower. A chi-squared test of independence was again performed in order to study the relationship between the language and the accuracy. The test showed no significant relationship between the two variables, $\chi^2(4, N = 1500) = 5.85, p = 0.21$.

To understand how good ChatGPT is at answering CSR questions, its accuracy in the task was compared to that of two other models on the same dataset. The first model, XLM-R (Conneau et al., 2020), is a large multilingual language model often used as a benchmark, while the second, MCP(XLM-R), is a variant of the first model which uses multilingual contrastive pre-training (MCP) (Lin et al., 2021). The results of the 3 models are shown in Table 3.4.

Differently from SA, ChatGPT performs overall

better than the two benchmark models in the CSR task. It has a higher accuracy in every language with the exception of Italian, for which the XLM-R model with MCP performs slightly better. These results suggest that the CSR task is harder than the SA task, with both ChatGPT and XLM-R performing worse, but also that ChatGPT is better at CSR than SA, at least compared to XLM-R.

In order to assess how much ChatGPT’s answers vary in this task, the standard deviation was computed as for the sentiment analysis task. The results (Table 3.3) show again that both values vary similarly in all languages. With regards to accuracy, English and Italian have the lowest deviation for accuracy while Spanish has the highest. On the other hand, German has the lowest deviation for confidence while English has the highest.

As a way to further analyse the calibration of ChatGPT, the ECE and MCE were computed for each language and are presented in Table 3.3.

The results are overall worse than in the SA task, with the lowest ECE being 16 for English and French and the highest being 21 for Italian. The analysis of the accuracy with the ECE seems to indicate that for the CSR task, English and Italian can be identified as the best and worst languages. The MCE results for this task are also overall higher than for SA, although once again heavily dependent on the small bins.

Lastly, it is important to note that, differently from the first task, in the CSR task English has both the best accuracy and the best ECE, suggesting a potential overall better performance.

4 Discussion

4.1 Performance in High-Resource Languages

With regard to ChatGPT’s accuracy in performing NLP tasks in high-resource languages, no significant difference was found between languages. Moreover, except for English having the highest accuracy in both tasks, the other languages’ performance did not match their respective resource levels. For example, despite having the least resources of the five, Italian has the best accuracy in the SA task together with English. This suggests that ChatGPT can perform NLP tasks equally well in all high-

Table 3.3: Summary of the results per language of ChatGPT in the common sense reasoning task. The values shown are mean accuracy and confidence (with standard deviation) as well as expected calibration error and maximum calibration error. There is no significant difference in accuracy between the languages.

Language	Accuracy \pm SD	Confidence \pm SD	ECE	MCE
English	70.0 \pm 1.5	85.4 \pm 2.2	15.8	60.0
French	65.3 \pm 1.5	80.7 \pm 2.0	15.7	40.0
Italian	60.6 \pm 1.5	82.3 \pm 1.1	21.8	60.0
Spanish	64.3 \pm 3.1	83.9 \pm 1.2	19.7	60.0
German	65.0 \pm 2.5	82.4 \pm 0.8	17.9	40.0

resource languages, regardless of the exact resource level.

Comparing the ECE and MCE results for each language in both tasks also shows no great disparity among the languages. No language has the best or worst ECE or MCE in both tasks, indicating that, among high-resource languages, a larger number of resources does not help ChatGPT to better predict the accuracy of its answers.

Comparing the results in the two tasks shows that ChatGPT has a harder time answering CSR questions than performing sentiment analysis. This is suggested by the higher accuracy and lower ECE in the latter. However, compared to other large language models such as XLM-R, ChatGPT seems to be more accurate in CSR and less in SA. This suggests that ChatGPT’s worse performance in CSR is due to the task being overall harder, and not to incapacity. Nonetheless, this does not explain why the average confidence in CSR is higher. On the other hand, it appears that ChatGPT is not as skilled in sentiment analysis compared to other state-of-the-art models. Moreover, analysing the standard deviation suggests that regardless of the task, Chat-

GPT’s answers and its confidence in them vary only slightly.

Lastly, English having the best accuracy in both tasks as well as the best ECE for the CSR task suggests that a significant difference in performance among high-resource languages might appear with more or harder tasks.

4.2 Confidence Calibration

The additional request to give a confidence value for every answer also led to interesting findings about ChatGPT. Firstly, the range of confidences: ChatGPT never gave a confidence value below 50%. This is a great issue as it points towards ChatGPT not being aware of its limits. When ChatGPT is faced with a complex task, it should be able to tell its users that its answers may be inaccurate. Instead, it seems that ChatGPT tends to give a confidence value that makes its answer seem trustworthy although it may not be.

Secondly, its calibration: across all languages and tasks, the average ECE is 15. This value can be considered acceptable, as it suggests ChatGPT has a general idea of its answers’ accuracy. However, when considered with the calibration plots, it becomes more worrisome. Although the ECE uses an absolute difference, Figures 3.1 and 3.4 show that the accuracy is mostly lower than the confidence, suggesting that ChatGPT is often overconfident in its abilities.

Overall, considering the range of confidences and their calibration, it seems that ChatGPT’s confidence in its answers cannot be trusted and should at best be considered as an approximate higher-bound estimate.

Table 3.4: Accuracy of ChatGPT in the common sense reasoning task compared to that of two benchmark models (XLM-R, MCP(XLM-R)) on the X-CSQA dataset. ChatGPT’s performance is better in all languages except Italian.

Language	ChatGPT	XLM-R	XLM-R _{MCP}
English	70.0	66.7	69.5
French	65.3	60.3	60.0
Italian	60.1	58.2	60.3
Spanish	64.3	59.5	61.4
German	65.0	56.1	59.3

4.3 Limitations

In the process of studying ChatGPT’s performance and awareness in different high-resource languages, some limitations were encountered. These limitations, which mainly concern the data, need to be acknowledged to gain a more comprehensive perspective of the results and to highlight areas of improvement for future research.

Firstly, the amount of data. As the data needed to be gathered manually, the amount of data used was quite limited. This caused multiple bins to have very small sizes and this may have influenced the results. With the OpenAI API now available, gathering larger amounts of data should be easier.

Secondly, the number of tasks. Although the tasks selected show similar results regarding the languages’ performances, using two tasks is not sufficient to generalise about ChatGPT’s abilities confidently. Repeating the experiment with more NLP tasks would potentially strengthen the conclusions.

Lastly, the translated data of the CSR tasks. As explained by the creators of the dataset (Lin et al., 2021), the data in non-English languages was translated using machine translation and its quality was ensured through an automatic evaluation. However, upon close inspection from native speakers, it appears that some questions or answers contain mistakes that hinder their clarity. This might have influenced ChatGPT to perform worse in non-English languages in the CSR task.

5 Conclusions

In this paper, two research questions regarding ChatGPT were defined as the goal of the research. Firstly, whether ChatGPT is equally accurate in high-resource languages, and secondly whether ChatGPT’s confidence in its answers is well-calibrated. In order to answer these two questions, ChatGPT was asked to perform 2 NLP tasks (sentiment analysis and common sense reasoning) in 5 high-resource languages (English, French, Italian, Spanish and German) and to give a confidence level for each of its answers.

The results showed that all the languages achieve similar accuracy in both tasks, with no significant difference between them, therefore providing a positive answer to the first research question. The range

of confidence levels given, together with the calibration plot and the expected calibration error, suggests that ChatGPT does not know when it lacks knowledge and that it is often overconfident. This answers negatively the second research question. ChatGPT is equally accurate in high-resource languages but its confidence in its answers is not well calibrated.

With regard to the performance in the two NLP tasks, the comparison to the XLM-R model suggests that ChatGPT is more suited for common sense reasoning than for sentiment analysis.

5.1 Future Research

This study provides valuable insights into ChatGPT’s multilingual abilities and confidence calibration. However, as new GPT models get released and gain more and more users, it is fundamental to keep studying their strengths and limitations. The findings of this paper can be used as a basis for a multitude of research areas.

The same experiment could be repeated on newer models, such as GPT4, to assess whether the performance between languages changes, whether they give a wider range of confidence levels or whether they have better confidence calibration.

Finally, more NLP tasks or high- to medium-resource languages could be added. Adding tasks of different degrees of difficulty would give the findings a broader generalisation and would confirm whether all high-resource languages perform equally or whether the performance depends on the task. Adding more languages would help assess the resource level at which languages start performing significantly worse than English.

References

- Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., . . . others (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.
- Barbieri, F., Espinosa Anke, L., & Camacho-Collados, J. (2022, June). XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the thirteenth language resources and evaluation confer-*

- ence (pp. 258–266). Marseille, France: European Language Resources Association.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., . . . others (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877–1901.
- Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.
- commoncrawl. (n.d.). *Statistics of Common Crawl Monthly Archives by commoncrawl*. Retrieved 01-04-2023, from <https://commoncrawl.github.io/cc-crawl-statistics/>
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., . . . Stoyanov, V. (2020, July). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8440–8451). Online: Association for Computational Linguistics.
- Davis, E., & Marcus, G. (2015, 8). Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of The ACM*, 58(9), 92–103.
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th international conference on machine learning - volume 70* (p. 1321–1330). JMLR.org.
- Jiao, W., Wang, W., Huang, J.-t., Wang, X., & Tu, Z. (2023). Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.
- Lin, B. Y., Lee, S., Qiao, X., & Ren, X. (2021). *Common sense beyond english: Evaluating and improving multilingual language models for commonsense reasoning*.
- Medhat, W., Hassan, A. E., & Korashy, H. (2014, 12). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093–1113.
- Petrosyan, A. (2023). *Most used languages online by share of websites 2023 — Statista*. Retrieved 01-04-2023, from <https://www.statista.com/statistics/262946/most-common-languages-on-the-internet/>
- Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., & Yang, D. (2023). Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.
- Ruby, D. (2023). *ChatGPT Statistics for 2023 (New Data + GPT-4 Facts)*. Retrieved 01-04-2023, from <https://www.demandsage.com/chatgpt-statistics/>

A Appendix

A.1 Decision Between Using English or Translated Prompts

In order to decide whether to use English prompts or translated prompts a small experiment was conducted. For each language, the first 30 data points of the sentiment analysis task were selected. ChatGPT was instructed to perform sentiment analysis both in English and in the language of the datapoints. The accuracy when using the English prompt or the translated prompts was then computed and can be found in Table A.1.

Table A.1: Accuracy in the first 30 prompts of the sentiment analysis task when using English or translated prompts. German shows a significant difference in accuracy between the two options.

Language	English	Translated
English	76.7	N/A
French	73.3	73.3
Italian	76.7	76.7
Spanish	56.7	66.7
German	56.7	70.0

Since German showed the greatest difference in accuracy, it was selected for a McNemar test of significance. The McNemar test determined that there was a statistically significant difference in the accuracy of ChatGPT between using an English or German prompt, $p = 0.03$. Since at least one language performed significantly better when using a translated prompt, it was decided to use translated prompts for all languages and tasks.

A.2 Translated Prompts

A.2.1 Sentiment Analysis

- **French:** Bonjour, j'aimerais que tu fasses une analyse de sentiment sur 15 phrases françaises. Classe chaque phrase comme "Positive" si elle exprime un sentiment positif, "Negative" si elle exprime un sentiment négatif ou "Neutral" si elle est neutre." Aussi, évalue ta confiance dans la réponse que tu as donnée entre 0 et 100 %. La réponse doit être une liste au format "Sentiment (confiance%)". Réponds "compris" si tu

as compris. La liste des phrases suivra.

- **Italian:** Ciao, vorrei che tu eseguiessi un'analisi del sentimento su 15 frasi italiane. Classifica ogni frase come "Positive" se esprime un sentimento positivo, "Negative" se esprime un sentimento negativo o "Neutral" se è neutra. Inoltre, valuta quanta fiducia hai nella risposta che hai dato tra 0 e 100%. La risposta dovrebbe essere un elenco con il formato "Sentimento (fiducia%)". Rispondi "capito" se hai capito. Seguirà l'elenco delle frasi.
- **Spanish:** Hola, me gustaría que realizaras un análisis de sentimiento de 15 frases en español. Clasifica cada frase como "Positive" si expresa un sentimiento positivo, "Negative" si expresa un sentimiento negativo o "Neutral" si es neutra. Además, puntúa tu confianza en la respuesta que has dado entre 0 y 100%. La respuesta debe ser una lista con el formato "Sentimiento (confianza%)". Contesta "entendido" si lo has entendido. A continuación aparecerá la lista de frases.
- **German:** Hallo, ich möchte Dich bitten, eine Stimmungsanalyse für 15 deutsche Sätze durchzuführen. Ordne jeden Satz als "Positive" ein, wenn er eine positive Stimmung hat, "Negative", wenn er eine negative Stimmung hat oder "Neutral", wenn er neutral ist. Bewerte bitte auch Dein Vertrauen in Deine Antwort zwischen 0 und 100%. Die Antwort muss eine Liste im Format "Stimmungsart (Vertrauen%)" sein. Antworte mit "verstanden", wenn Du verstanden hast. Es folgt eine Liste von Sätzen.

A.2.2 Common Sense Reasoning

- **French:** Bonjour, j'aimerais que tu répondes à 15 questions de bon sens à choix multiple en français. Pour chaque question, donne la lettre qui correspond à la bonne réponse. Aussi, évalue ta confiance dans la réponse que tu as donnée entre 0 et 100 %. La réponse doit être une liste au format "Lettre (confiance%)". Réponds "compris" si tu as compris. La liste des questions suivra.
- **Italian:** Ciao, vorrei che tu rispondessi a 15 domande di senso comune a scelta mul-

tipla in italiano. Per ogni domanda, scegli la lettera che corrisponde alla risposta corretta. Inoltre, valuta quanta fiducia hai nella risposta che hai dato tra 0 e 100%. La risposta dovrebbe essere un elenco con il formato "Lettera (fiducia%)". Rispondi "capito" se hai capito. Seguirà l'elenco delle domande.

- **Spanish:** Hola, me gustaría que respondieras a 15 preguntas de sentido común de opción múltiple en español. Para cada pregunta, por favor, indique la letra que corresponde a la respuesta correcta. Además, puntúa tu confianza en la respuesta que has dado entre 0 y 100%. La respuesta debe ser una lista con el formato "Letra (confianza%)". Contesta "entendido" si lo has entendido. A continuación aparecerá la lista de preguntas.
- **German:** Hallo, ich möchte Dich bitten, 15 Multiple-Choice-Fragen des gesunden Menschenverstands auf Deutsch zu beantworten. Bitte gib für jede Frage den Buchstaben an, der der richtigen Antwort entspricht. Bewerte bitte auch Dein Vertrauen in Deine Antwort zwischen 0 und 100%. Die Antwort muss eine Liste im Format "Buchstabe (Vertrauen%)" sein. Antworte mit "verstanden", wenn Du verstanden hast. Es folgt die Liste von Fragen.

A.3 Sample ChatGPT Queries and Responses in Both Tasks

In this section, for each language, six examples of ChatGPT performing sentiment analysis and common sense reasoning are presented. For each language and task, three correct and three incorrect answers were randomly sampled. The tables present the tweet or question (and options) ChatGPT was given, ChatGPT's answer with its confidence level as well as the correct answer. In every table, the first three rows present correct answers and the last three present wrong answers.

Table A.2: Three correct and three incorrect answers ChatGPT gave while performing sentiment analysis in English.

Tweet	Answer	Correct
@user on Comedy Central tonight @ 9...I'm a Grown Little Man and Seriously Funny...I'll be glued to the tv	Positive (80%)	Positive
"Yo, if you paid \$57-\$77 to go see Eric Church at the State Fair...may the Lord have mercy on your souls. "	Negative (80%)	Negative
"So who's next? Maduro? A little young, perhaps, but it is Venezuela.... "	Neutral (75%)	Neutral
looks like I'm done with Kpop guys	Negative (89%)	Neutral
If I'm Dustin Johnson I wouldn't care about my Sunday meltdowns going home to Paulina. #sheesh	Positive (80%)	Negative
I always thought Prince was the 18th best songwriter of all time but it's nice of Rolling Stone to confirm that.	Neutral (75%)	Positive

Table A.3: Three correct and three incorrect answers ChatGPT gave while performing sentiment analysis in French.

Tweet	Answer	Correct
"Sévère réquisitoire de l'ex-ministre de l'Écologie, Delphine #Batho, contre l'action du chef de l'État dans un... http"	Negative (90%)	Negative
Formation aux rencontres écologiques d'été avec un futur eurodéputé espagnol issu des Indignés. Passionnant! #ree14 http	Positive (75%)	Positive
Fédération des Parcs Naturels Régionaux — @user http	Neutral (65%)	Neutral
Aides à l'agriculture biologique : les écologistes mobilisés http	Neutral (70%)	Positive
Région PACA : le nouvel écosystème de l'export en ordre de marche http	Positive (75%)	Neutral
Si on ne peut même plus rêver... #animaux #amitié http	Neutral (65%)	Negative

Table A.4: Three correct and three incorrect answers ChatGPT gave while performing sentiment analysis in Italian.

Tweet	Answer	Correct
"Buongiorno e ""buona scuola"" a tutti... finalmente è proprio il caso di dirlo!"	Positive (80%)	Positive
Dimettiti incapace! @user http	Negative (70%)	Negative
Lettera al premier mario monti Corriere della Sera Inviato da iPad http	Neutral (60%)	Neutral
"Metti un ""like"" per la #scuola paritaria #iniziativa http #labuonascuola"	Positive (80%)	Neutral
"@user si certo, ma questa definizione passa spesso in secondo piano a seconda di chi viene indagato"	Neutral (70%)	Negative
Batteria 2%...buonanotte a tutti	Negative (90%)	Neutral

Table A.5: Three correct and three incorrect answers ChatGPT gave while performing sentiment analysis in Spanish.

Tweet	Answer	Correct
"@user hombreee, poco a poco, los 5 primeros minutos que montes en bici te sabrán a gloria así que ánimo que todo llega"	Positive (90%)	Positive
@user pero eso no es gracioso	Negative (80%)	Negative
"A veces cuando leo sus tuits y no entiendo algo, me dan ganas de preguntarles para entender mejor. Pero soy tímido y me quedo con la duda"	Neutral (75%)	Neutral
Debo dejar de hacer bromas porque cuando hable en serio no me creerán	Negative (70%)	Neutral
Bienvenida a los juegos mentales	Negative (80%)	Positive
Esta para unos buenos mates y galletitas	Positive (90%)	Neutral

Table A.6: Three correct and three incorrect answers ChatGPT gave while performing sentiment analysis in German.

Tweet	Answer	Correct
"@user Weitermachen, hopp hopp. Steht noch genug aus ;)"	Positive (80%)	Positive
"RT @user: Entrüstung über Merkel-Abhöraktion. Berechtigt? Ja, natürlich! Ja!! Aber Front im Cyberwar verläuft längst woanders http"	Negative (75%)	Negative
RT @user: Sei nicht zu müde! Nächster Halt: Romantik mit #Flirtbotschaft http #München	Neutral (70%)	Neutral
Ich habe ein @user-Video positiv bewertet: http Let's (schlechte) Parodie: 24SpeckTV	Neutral (50%)	Positive
@user große katastrophe :P	Negative (65%)	Positive
"Gute Besserung mutee RT ""@user: Endlich muss ich zum Arzt gehen :—"""	Positive (75%)	Negative

Table A.7: Three correct and three incorrect answers ChatGPT gave while performing common sense reasoning in English.

Question	Answer	Correct
What castle is built upon Castle Rock in Scotland? A: capturing pawn. B: london. C: edinburgh. D: germany. E: europe.	C (90%)	C
John didn't like eating the hamburger. It made him feel what? A: enjoy. B: nausea. C: sad. D: death. E: satisfaction.	B (95%)	B
Everyone is special. Everyone has what? A: feelings. B: values. C: unique personality. D: experiences. E: different standards.	C (95%)	C
Some people really loved beer, they call it what gold? A: alcohol in. B: liquid. C: intoxicating. D: hair of the dog. E: harmful.	D (70%)	B
38. Stopping being married to her allowed him to again pursue his dreams, it brought him what? A: depression. B: wrong. C: relief. D: remembrance. E: pleasure.	C (80%)	E
After the guy is successful in cashing in his check, what does he feel? A: sad. B: quitting. C: extra money. D: leave. E: great joy.	C (85%)	E

Table A.8: Three correct and incorrect answers ChatGPT gave while performing common sense reasoning in French.

Question	Answer	Correct
Où les élèves collent-ils des chewing-gums à l'école? A: livres. B: trottoir. C: fontaine d'eau. D: étagère. E: films.	C (80%)	C
Sauter par-dessus une corde est une forme de quoi ? A: maux de tête. B: fun. C: se réchauffer. D: exercice. E: brisé.	D (90%)	D
Où les étudiants sont-ils susceptibles d'utiliser une structure de stationnement ? A: grande ville. B: centre commercial. C: campus universitaire. D: derrière le garage.. E: chicago.	C (75%)	C
Que doit faire un enseignant pour ses élèves ? A: une meilleure connaissance. B: groupe d'étudiants. C: préparer le déjeuner. D: test de temps. E: énoncer les faits..	A (85%)	E
Quel est l'endroit idéal pour avoir un chat ? A: chaise confortable. B: walmart. C: société humaine. D: étage. E: appui de fenêtre.	E (80%)	C
Où peut-on voir un point de repère entouré de plusieurs kilomètres de plantes, si ce n'est dans un parc ? A: campagne. B: parc national. C: désert. D: carte. E: ville.	B (80%)	A

Table A.9: Three correct and three incorrect answers ChatGPT gave while performing common sense reasoning in Italian.

Question	Answer	Correct
Qual è il solito motivo per cui la gente si diverte a giocare? A: felicità. B: intrattenimento. C: competitività. D: rabbia. E: vincente.	B (90%)	B
Gli esseri umani imparano l'attualità da quale articolo di stampa? A: domandare l'autorità. B: leggere i giornali. C: sciogliere. D: agire. E: dizionario.	B (80%)	B
Dove si sentirebbe più libero un furetto? A: all'aperto. B: carolina del nord. C: nord america. D: gran bretagna. E: aula.	A (90%)	A
Dove sarebbe improbabile vedere delle impronte a causa dell'assenza di persone? A: terra. B: tappeto. C: camino. D: tetto. E: superficie di terra.	A (85%)	D
Aveva una relazione, cosa è finito con la scoperta del marito? A: relazione. B: matrimonio. C: fedeltà. D: amore. E: divorzio.	E (75%)	B
Sarah ama dare assistenza. Non chiede nulla in cambio. A cosa probabilmente Sarah tiene di più? A: i suoi capelli. B: circostanze migliori. C: felicità. D: imparare nuove competenze. E: persone aiutate.	C (85%)	E

Table A.10: Three correct and three incorrect answers ChatGPT gave while performing common sense reasoning in Spanish.

Question	Answer	Correct
¿Dónde podrías averiguar sobre una vaca que no existe? A: al aire libre. B: libro de cuentos. C: campo. D: Tienda de juguetes. E: corral.	B (90%)	B
El empleado preguntó qué tipo de carne y pan quería el cliente, y eligieron tener un qué? A: bifurcación. B: congelador. C: horno. D: sartén. E: sándwich de jamón.	E (80%)	E
¿Qué es lo que Dios nunca hará según la religión? A: cualquier cosa. B: hacer milagros. C: juzgar a la gente. D: Juez de hombres. E: en todas partes.	C (85%)	C
¿Qué hará un agente inmobiliario sin escrúpulos para que un listado parezca más atractivo? A: contratar una banda. B: exagerar. C: encantador. D: engañar. E: mentira.	B (70%)	E
Estaba encontrando información a través de la meditación y el yoga, ¿qué estaba buscando? A: obtener respuestas. B: úlceras. C: poder. D: felicidad. E: respeto.	A (90%)	D
Kuwait y otros países de la región no siempre se asocian inmediatamente con ella, pero están situados en donde? A: estados del golfo. B: oriente medio. C: arabia. D: kuwait. E: asia	B (85%)	E

Table A.11: Three correct and three incorrect answers ChatGPT gave while performing common sense reasoning in German.

Question	Answer	Correct
Ein Arzt kann ein Antibiotikum in vielen Formen verschreiben, welcher Typ könnte eine Gel-Beschichtung haben? A: Kapsel. B: Medizinschrank. C: bakterielle Infektion heilen. D: vorgeschrieben. E: Apotheke.	A (80%)	A
Wo kann man einen Tennisplatz für Nicht-Profis finden? A: Erholungszentrum. B: Country-Club. C: Sportverein. D: College-Campus. E: zoo.	A (90%)	A
Welcher Teil eines Grundstücks darf neben einer Einfahrt liegen? A: Auto weiterfahren. B: Nachbarschaft. C: Hof. D: Unterteilung. E: Vorort.	C (85%)	C
Warum sollte eine Person an ihrem Arbeitsplatz bleiben, ohne bezahlt zu werden? A: Gott sei Dank. B: Schlange stehen. C: gerne arbeiten. D: Hilfe anbieten. E: Freiwillige.	E (70%)	C
Wo werden Sie wahrscheinlich Sterne aus dem Fenster sehen? A: Spaceshuttle. B: Gebäude. C: auf dem Rasen. D: Haus eines Freundes. E: Wand.	D (75%)	A
Warum verdienen Sie kein Geld, während Sie zu Mittag essen? A: die Arbeit einstellen. B: Nahrung zu sich nehmen. C: Essen besorgen. D: Nahrung finden. E: zu voll.	B (90%)	A