*Journal of Intelligence Studies in Business*
Vol. 14 No. 2 (2024) pp. 77–97
Open Access: Freely available at: https://jisib.com/
https://doi.org/10.37380/jisib.v14.i2.2547

# Comparison of effectiveness between ChatGPT 3.5 and 4 in understanding different natural languages

**Bernhard Erös***
*Technical Department, University of Applied Sciences FH Campus Vienna*
*Vienna, Austria*
*0009-0001-9584-8789*
*eroes.bernhard@gmail.com*

**Christoph Gritsch**
*Doctoral School of Applied Informatics and Applied Mathematics, Obuda University*
*Budapest, Hungary*
*0009-0000-8874-9427*

**Andrea Tick**
*Keleti Karoly Faculty of Business and Management, Obuda University*
*Budapest, Hungary,*
*0000-0002-3139-6509*

**Philipp Rosenberger**
*Technical Department, University of Applied Sciences FH Campus Vienna*
*Vienna, Austria*
*0000-0001-8157-111X*

**ABSTRACT:** This paper addresses the multilingual language understanding of ChatGPT-3.5 and 4 to investigate their performance with respect to languages with different degrees of prevalence on the internet. ChatGPT's training data mostly consists of website content. As the language distribution is unevenly allocated and a low number of languages is used on websites this should impact performance.
Both ChatGPT versions should rate reviews between 1 to 5 stars based solely on the product description and the review texts. Therefore, 500 e-commerce reviews are collected for each of five languages: English, German, Dutch, Korean and Hindi, which are evenly distributed at 100 reviews per star rating. The evaluation methods and metrics used in this study include t-tests, confusion matrices, macro F1 values and a defined cumulative star deviation.
The results indicate a significant correlation between the degree of dissemination and the accuracy of the ChatGPT-3.5 evaluation. In direct comparison, ChatGPT-4 shows superior accuracy in all languages studied, while maintaining acceptable performance in less represented languages. The hypothesis that ChatGPT-4 scoring accuracy increases with an increase in the number of words in reviews in less represented languages could not be confirmed. These findings illustrate the influence of the selected language on the interaction with ChatGPT and its language comprehension, which suggests that multilingualism should be given greater consideration in the future development and optimization of large language models.

**KEYWORDS:** ChatGPT, E-Commerce, Generative Pretrained Transformer, Large Language Model, Natural Language Understanding

* Corresponding Author

## 1. INTRODUCTION

The increasing popularity and performance of artificial intelligence (AI) makes it suitable for a wide range of tasks. As AI continues to evolve, its potential to influence various applications increases (Cekuls 2023). One common application of such artificial intelligence are chatbots. Therefore, this research focuses on the performance of such chatbots in understanding and interpreting multilingual content and its context. The AI chatbot ChatGPT (GPT) form OpenAI is employed for this purpose (Kantrowitz 2024). Such chatbots are based on large language models (LLM) (Alec Radford and Karthik Narasimhan 2018). In addition to social media, which are examined in a multilingual context based on region and culture, e.g. (Das et al. 2023; Leong et al. 2023), the e-commerce sector is selected as a multilingual area. This is the starting point for this study, which examines the common practice of reviews and a depiction of customer satisfaction or sentiment towards the respective product in the form of star ratings in different languages in connection with GPT.

According to OpenAI, their models were trained with the following data: texts, books, news articles, scientific articles and also websites (Zhu and Luo 2022). As websites form part of the training data for these advanced transformer models (TFM) (Zhao et al. 2023), the languages in which websites are written were examined. This showed that only a few specific languages are used to create websites (Web Technology Surveys 2024). ChatGPT-3.5, which was available free of charge at the time of writing, has 175 billion parameters, while the paid version 4 has over 1 trillion parameters and thus increased performance even further (Achiam et al. 2023).

In relation to natural language understanding (NLU), especially in the multilingual field, several studies and investigations have been carried out with LLM (Ahuja et al. 2023; Zhao et al. 2024). GPT was often compared with other models in order to draw direct performance comparisons (Naveed et al. 2023). This study follows the approach of investigating assumptions about the effect of the language used and the length of reviews on the language comprehension of GPT.

In this study, the language comprehension of the AI chatbot in version 3.5 as well as with version 4 will be tested. Specifically, this means whether GPT correctly understands and interprets the context and meaning of input in different natural languages. Five languages with different levels of representation on the internet are examined for this purpose. Product reviews from e-commerce platforms and the corresponding star ratings in the selected languages serve as data and metrics. GPT should classify these reviews correctly, i.e. recognize as precisely as possible a very negative to very positive sentiment about a product (depicted as 1-5 stars) in order to rate it accordingly from 1 to 5. The following research questions arise.

**RQ:** How does language choice influence the effectiveness of ChatGPT in recognizing the sentiment of review texts in different languages?

**RSubQ1**: Are more represented languages more accurately understood or recognized by ChatGPT-3.5 than less represented languages?

**RSubQ2**: Which of the two models, ChatGPT-3.5 or 4, is more accurate in recognizing the sentiment of different natural languages?

**RSubQ3**: For less represented languages, does more context result in a more accurate assessment of ChatGPT-4?

Within this paper, the current literature is reviewed regarding commonly used transformers and large language models and their accuracy. This is followed by the hypotheses and design of our study where the used statistics are also explained in greater detail. Afterwards the results are presented and discussed. Lastly, a conclusion and critical reflection of the limitations is given

## 2. PRELIMINARY LITETARURE REVIEW

Transformer models were developed to enable a more efficient translation of natural language and are also referred to as pretrained language models (Vaswani et al. 2017). The spread and establishment of these models has led to significant breakthroughs in the machine processing of language. They led to the further development of large language models, which refer to the fact that natural language can be represented as a language model (Russell and Norvig 2022). The first LLM was T5 (Text-to-Text-Transformer) (Naveed et al. 2023). The abbreviation GPT stands for Generative Pretrained Transformer and was introduced by OpenAI (Alec Radford and Karthik Narasimhan 2018). The structure of GPT-1 is based on the decoder

block of the encoder-decoder model, and the model specifications also largely correspond to the Transformer model originally introduced by Google (Devlin et al. 2018).

For efficient text processing of such models, texts are broken down into tokens (Bishop and Bishop 2024). Relationships in syntax and semantics are essential for the generalization of natural language and are mapped via embeddings. A distinction is made between two types of embedding: word embeddings and positional embeddings. Word embeddings use dense high-dimensional vectors, whereby related words have similar vectors (Russell and Norvig 2022; Bishop and Bishop 2024). Deep feed forward neural networks with a large number of layers are components of TFM as well as LLM (Vaswani et al. 2017; Alec Radford et al. 2019). Another central component of transformers is a self-attention mechanism that identifies relevant input content (Vaswani et al. 2017). In several steps of this algorithm, an attention score is formed by including query vectors, key vectors and value vectors (Paaß and Giesselbach 2023). These values are normalized and a softmax function is used to generate probability values for all tokens that add up to 1 in order to ultimately generate a self-attention embedding (Hastie et al. 2017).

Recurrent neural networks as well as long-short-term memory networks use supervised learning as a learning method (Schmidhuber 2015). The BERT (Bidirectional Encoder Representations from Transformers) transformer uses two approaches of the unsupervised learning method, namely masked language modeling and sentence prediction (Devlin et al. 2018; Hung Vo et al. 2025; Osváth et al. 2023). GPT uses semi-supervised learning as a combination of both learning methods in the form of autoregressive language modeling (Alec Radford et al. 2019). This is a purely statistical language model which states that language follows a natural sequential order. These models, therefore, calculate the probability of the next word based on the previous text sequence (Alec Radford et al. 2019).

$$p(x) = \prod_{i=1}^{n} p(s_n | s_1, s_2, \dots, s_{n-1}) \quad (1)$$

While statistical methods predict sequential data, reinforcement learning can further enhance the training process (Paaß and Giesselbach 2023). The goal of reinforcement learning is that a model chooses the right actions in order to maximize its reward. The reward for a model is based on a corresponding reward function (Paaß and Giesselbach 2023). GPT, on the other hand, has no access to such a reward function and must be trained by incorporating human feedback (reinforcement learning from human feedback) in order to approximate it (Kaufmann et al. 2023). Further refinement for specific NLP tasks is achieved with fine-tuning (Paaß and Giesselbach 2023). Important predefined hyperparameters for fine-tuning determine a model's performance: learning rate, epoch, batch size (Jin et al. 2023; Smith et al. 2017). To reduce model overfitting techniques such as dropout are employed (Bishop and Bishop 2024; Zhang and Bottou 2024).

Additionally, mini-batch training is used to reduce vanishing gradients, which are a hindrance when determining a minimum of the loss function, leading to better overall performance and stability of a model (Paaß and Giesselbach 2023).

LLM are state-of-the-art systems that can both process and generate texts with human-like performance (Naveed et al. 2023). They can use transfer learning to apply knowledge from one domain or task to other, related domains or tasks, given a sufficiently large amount of data (Bishop and Bishop 2024; Kalyan 2024). Reinforcement learning from human feedback is used to align these LLM, which learn on the basis of human intentions and values, the defined "HHH" criteria – "HHH" stands for helpful, honest and harmless (Naveed et al. 2023).

GPT-1 has a total of 117 million parameters. For pre-training, the Bookcorpus dataset was used in a cleaned form, i.e. with standardized punctuation and the removal of superfluous spaces (Zhu et al. 2015). Input transformations are used to fine-tune this model (Alec Radford and Karthik Narasimhan 2018). The currently freely available version of GPT-3.5 is based on the original GPT-3 architecture from OpenAI, which was introduced as a further development of GPT-1 and GPT-2 (Brown et al. 2020). GPT-3.5 has 175 billion parameters (Brown et al. 2020). Gradient clipping, which accelerates the convergence of the training process, was also used in its pre-training (Zhang et al. 2019). Its training data mainly consists of filtered CommonCrawl data with 410 billion tokens (60%) and WebText2 with 19 billion tokens (22%) (Brown et al. 2020).

For the most powerful model at the time of writing, GPT-4, OpenAI does not provide any detailed information regarding its architecture (exact number of parameters, hidden

layers, or training data mix) for competitive and security-related reasons. Over 1 trillion parameters and 8,192 or rather 32,768 tokens for inputs can be estimated from various sources (Achiam et al. 2023). GPT-4 incorporates multimodality approaches, i.e. prompts can also contain images or documents. During GPT-4's development, the focus was placed on a scalable loss function within its deep learning stack, as model-specific fine-tuning is not economically viable (Achiam et al. 2023). The loss of an autoregressive model depends on the model size N, the computational budget C and the data set D (Henighan et al. 2020). The loss is the sum of a non-reducible loss and a scalable power term which describes the Kullback-Leibner divergence, a measure of the difference between two probability functions (Bishop and Bishop 2024; Zhao et al. 2023). Tests with smaller, downscaled models showed the optimal model size to be proportional to the computational budget (Henighan et al. 2020):

$$N_{opt} \propto C^{0,7}$$

The safe use of GPT-4 was improved by means of "red teaming" in order to avoid harmful advice and incorrect information (Brundage et al. 2020). For this purpose, the reinforcement learning from human feedback algorithm was extended with a rule-based reward model, which evaluates inputs based on rules in order to further align the model with the HHH criteria (Glaese et al. 2022). OpenAI carried out performance comparisons in various languages covering 57 subject areas, as well as questions on morality and justice (Hendrycks et al. 2020b; Hendrycks et al. 2020a). The results in Fig. 1 show that performance varies depending on the language, with performance being highest in English. The languages observed and examined in this research are also presented in Fig. 1.

With the successor GPT-4o, OpenAI indicates a slight increase in performance for multilingual tasks, but without further details (Hello GPT-4o 2024). Despite improvements, GPT-4 still generates unreliable content, but performs 20% better than GPT-3.5 in internal evaluations (Achiam et al. 2023).

## 3. RESEARCH METHODOLOGY

During the research English, German, Dutch, Korean and Hindi were selected with different current levels of representation on the internet as shown in Fig. 2.

Based on the literature review, the research questions as well as after having selected the languages hypotheses were formulated and the research design has been developed. The statistical analysis in this work as well as the representation of the data was carried out using MS Excel.

### *A. Hypotheses*

After presenting the problem statement and the derived research question, as well as an introduction to the technical-theoretical background of this work, the following hypotheses are formulated.

The first hypothesis (**A1**) deals with the accuracy depending on the language used.

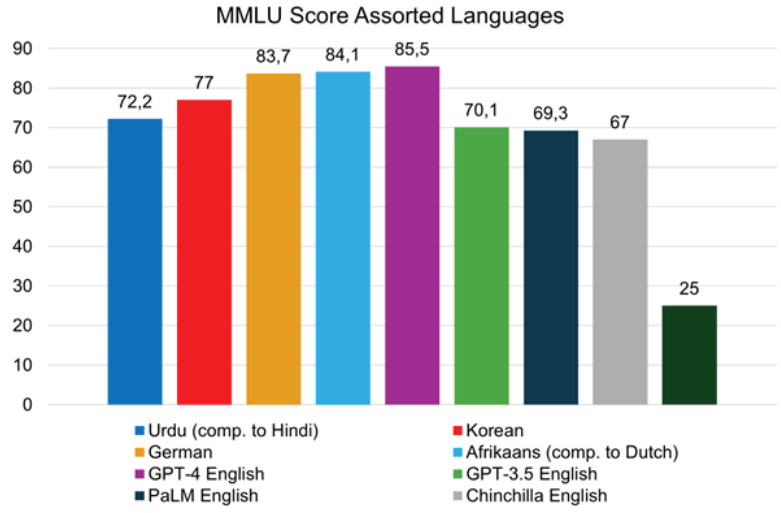


*Figure 1.* Accuracy of GPT–4 and GPT-3.5 for multilingual tasks
Note. based on Achiam et al. 2023.

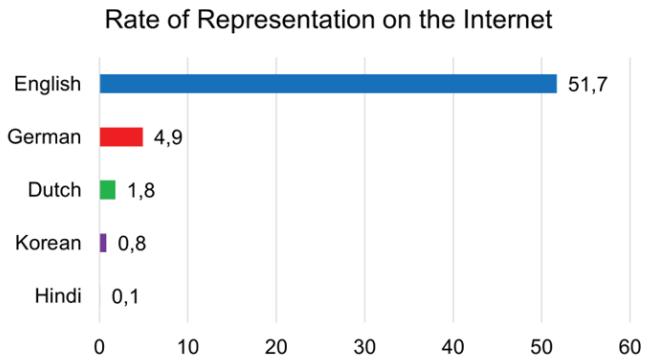


*Figure 2.* Content languages for websites
Note. based on Web Technology Surveys 2024.

- ChatGPT-3.5 understands languages represented more frequently on the internet more effectively and rates them more accurately than less represented languages.

As second part, ChatGPT-3.5 and ChatGPT-4 were compared. Therefore, within the hypothesis (**A2**) it was proposed that

- the performance of ChatGPT-4 in terms of language comprehension in different natural languages improves compared to ChatGPT-3.5.

As third hypothesis (**A3**), the increased accuracy with higher context and length in unrepresented languages was examined.

- It was assumed that more content or context in a less represented language leads to a more precise result when assessed by ChatGPT-4.

### *B. Research Study Design*

The research approach and the research flow is presented in Fig. 3. First, freely accessible reviews of e-commerce platforms have been collected in five languages, then they have been checked for validity in a KNIME workflow and their word counts were determined. Then these reviews have been reformulated into zero-shot prompts and submitted to the two ChatGPT models. The results were statistically analyzed to evaluate their performance across the selected languages. Finally, confusion matrices were created for GPT-3.5 and GPT-4, which show the values for accuracy, precision, recall or sensitivity and the relevant macro F1 value (average value of all F1 values of the five-star categories) for each star category.

Several hypothesis tests (were used during the evaluation process. To ensure adequate test quality, a statistic power of (have been chosen. In order to ensure statistic power and quality, the sample size of the reviews in the respective languages had to be large enough (Otte et al. 2018). The following approximate formula was used to determine a minimum sample size (Lehr 1992) with s representing standard deviation and d the minimum detectable difference.

$$n = \frac{16*s^2}{d^2} = \frac{16*0{,}5^2}{0{,}1^2} = \frac{16*0{,}25}{0{,}01} = 400 \qquad (2)$$

Since the formula is only an approximation and the selected sample size of 500 is 25% larger than the calculated value, the required statistic power can be ensured.

The observed and evaluated reviews were collected *ad hoc* on amazon.co.uk, amazon.de, amazon.nl, coupang.com and flipkart.com and were transferred without preprocessing as for example without the removal of stop words or spelling mistakes. The reviews and ratings are accessible at researchgate.net (Erös 2024). The reviews were collected across 16 different product categories, with books, films and music being excluded from the outset in order to avoid bias. Within the star categories, a variation between short and long reviews is present, which is why the word count of all 2,500 reviews is determined. Fig. 4 illustrates the mean word count of the cumulative reviews for each language.

Table 1 presents the minimum, maximum and mean values for each star category respectively. As this data represents an occasional sample, it is also checked whether the collected reviews reflect negative reviews (1 + 2 stars) and positive reviews (4 + 5 stars) as well. The classification is done using a multilingual BERT model in a two-class sentiment analysis (Github 2024b). The sentiment analysis in the KNIME workflow with the multilingual BERT model confirms the validity of the collected reviews with an accuracy equaling 0.79, and Cohen's Kappa being 0.58, which values are robust and comparable to the results in

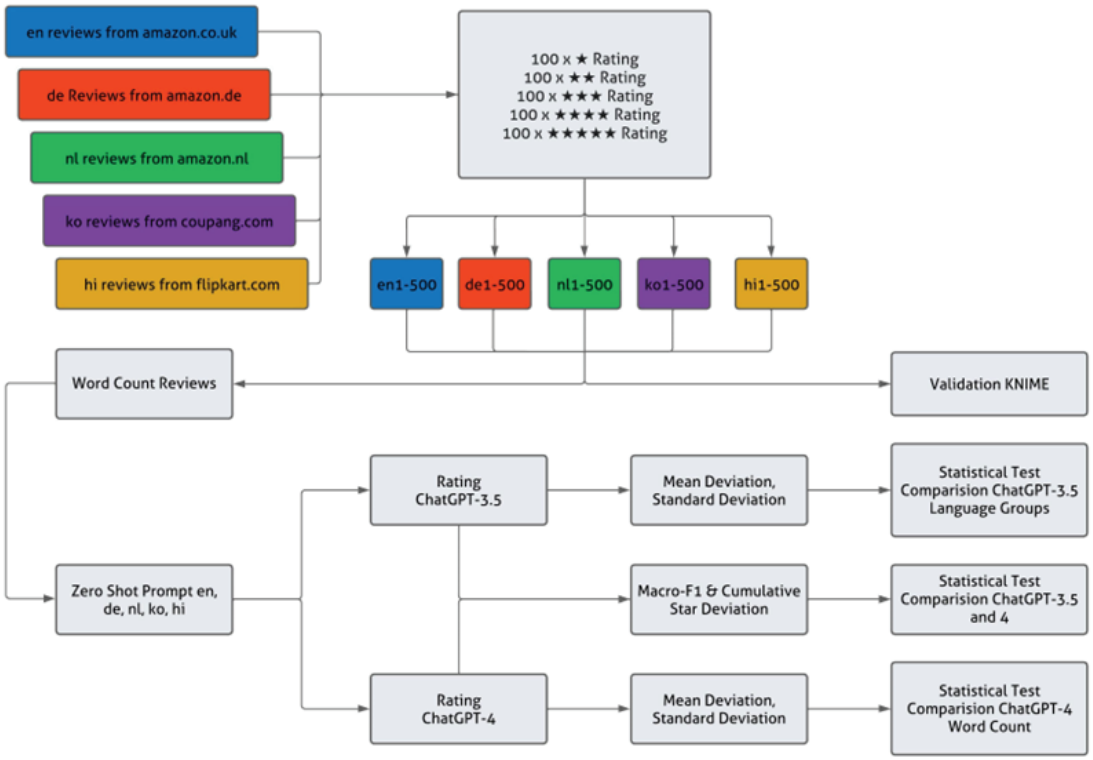


*Figure 3.* Study design

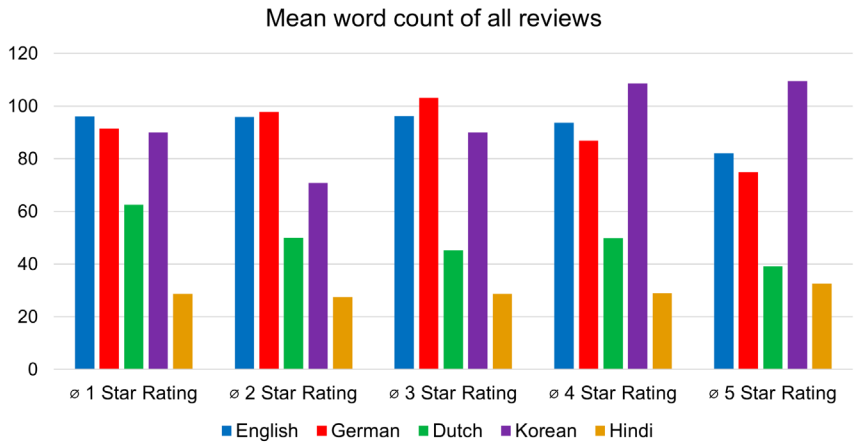


*Figure 4.* Graphical illustration of the mean word count of all reviews
Note. ⌀: average star rating of the respective star categories

*Table 1.* Minimum, maximum, and average word count of all reviews

| Language | Min. 1 Star | Max. 1 Star | ⌀ 1 Star | Min. 2 Stars | Max. 2 Stars | ⌀ 2 Stars | Min. 3 Stars | Max. 3 Stars | ⌀ 3 Stars | Min. 4 Stars | Max. 4 Stars | ⌀ 4 Stars | Min. 5 Stars | Max. 5 Stars | ⌀ 5 Stars |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| English | 8 | 451 | 96 | 11 | 582 | 96 | 13 | 483 | 96 | 13 | 437 | 94 | 5 | 793 | 82 |
| German | 3 | 320 | 91 | 6 | 852 | 98 | 8 | 481 | 103 | 7 | 261 | 87 | 4 | 210 | 75 |
| Dutch | 5 | 526 | 63 | 8 | 539 | 50 | 5 | 232 | 45 | 3 | 314 | 50 | 4 | 205 | 39 |
| Korean | 8 | 400 | 90 | 9 | 250 | 71 | 16 | 498 | 90 | 17 | 504 | 109 | 27 | 490 | 109 |
| Hindi | 3 | 83 | 29 | 2 | 89 | 28 | 2 | 79 | 29 | 2 | 76 | 29 | 5 | 86 | 33 |

*Note. ⌀: average star rating of the respective star category*

other studies (Das and Pedersen 2024; Pota et al. 2021)

A Cohen's Kappa value of is rated good (Döring 2023). A classification with three classes (negative, neutral, positive) would achieve lower values (Talaat 2023; Hendrycks et al. 2020b). All ratings of the reviews by ChatGPT were exported and are also available at researchgate.net (Erös 2024).

Before statistical comparisons, all product descriptions and reviews are formulated in zero-shot prompts and submitted to both ChatGPT models via the chat.openai.com web interface, with a total of 2,500 prompts per model being evaluated. Only integer star ratings (1–5) are expected as outputs from ChatGPT based on their understanding of the product description and associated review.

These outputs are statistically analyzed based on the mean deviation, standard deviation and the F1 value. In addition, a cumulative star deviation (CSD) has been developed for this research specifically:

$$CSD = \sum_{k=1}^{500} |LNG_k - LNGGPT_k| \quad (3)$$

It is therefore calculated from the sum of the absolute differences between a language, i.e. the star rating with a running identifier ($LGN_k$) and $LNGGPT_k$, the rating by one of the two GPT models.

The CSD measures the sum of the absolute differences between the star ratings and the outputs of both GPT models. In the case of deviations caused by ChatGPT, such as hallucinations, a prompt to the model is not repeated. In this case, a maximum possible deviation to the original star rating is included in the rating. On average, the possible maximum deviation corresponds to a deviation of 2.05 stars for all possible combinations of maximum deviations.

The next section presents the results of the statistical tests and the comparison of the accuracy or hallucination of the two selected ChatGPT versions in the case of the five selected languages, English, German, Dutch, Korean and Hindi.

## 4. RESULTS AND FINDINGS

The calculated statistical values for both GPT versions are shown in the following evaluation tables for each language. The notation is the following:

ø: average star rating of the respective star category
Δ: average deviation of the respective star category
s: standard deviation of the respective star category

The values highlighted in bold depict the higher or better results in each case.

### *C. Results for English*

Table 2 – Table 5 present the statistic evaluation (Table 2), the confusion matrix (Table 3), the rating performance comparison (Table 4) and the overall comparison of the two ChatGPT models in the case of English (Table 5).

According to the results in Table 2, 1-star ratings tend to be recognized as worse than the remaining star ratings by both models. With regard to the standard deviation, it can be seen that the dispersion is roughly equally distributed across all categories. Furthermore, although the mean deviation of GPT-4 in the 3 to 5 stars categories is marginally worse than with GPT-3.5, it is still two percentage points

*Table 2.* Evaluation English language GPT–3.5 and GPT–4

| | **GPT–3.5 ø** | **GPT–4 ø** | **GPT–3.5 Δ** | **GPT–4 Δ** | **GPT–3.5 s** | **GPT–4 s** | **GPT–3.5 CSD** | **GPT–4 CSD** |
|---|---|---|---|---|---|---|---|---|
| ★ | 1.81 | **1.65** | 0.81 | **0.65** | **0.506** | 0.539 | 213 | **208** |
| ★★ | 2.41 | **2.32** | 0.41 | **0.32** | 0.514 | **0.510** | | |
| ★★★ | **3.02** | 2.89 | **0.34** | 0.37 | **0.476** | 0.485 | | |
| ★★★★ | **3.81** | 3.69 | **0.27** | 0.37 | **0.446** | 0.562 | | |
| ★★★★★ | **4.70** | 4.67 | **0.30** | 0.33 | **0.461** | 0.493 | | |
| Mean | – | – | 0.43 | **0.41** | **0.481** | 0.518 | | |

more precise across all categories, which can also be seen in the confusion matrix below (Table 3). Moreover, the recall tends to be better for GPT-4, since only the 4-star category was slightly better in GPT-3.5 (Table 4).

Besides precision and recall, Table 4 presents the F1 values of the corresponding confusion matrix. It is shown that in the 1 to 3 stars categories when switching to GPT-4 the F1 values are higher. The overall accuracy increased by 2.6% percent when switching to GPT-4 (Table 5).

### *D. Results for German*

Table 6 – Table 9 display the statistic evaluation (Table 6), the confusion matrix (Table 7), the rating performance comparison (Table 8) and the overall comparison of the two ChatGPT models in the case of German (Table 9). In this case a reduced performance of both models for 1-star ratings can also be observed.

In Table 6, one of the most outstanding values was observed for GPT-3.5 (highlighted in blue). Besides the aforementioned overall reduced performance, GPT-3.5 in particular has

*Table 3.* Confusion Matrix for English language GPT–3.5/4

| | | Rating | | | | | Recall |
|---|---|---|---|---|---|---|---|
| | | ★ | ★★ | ★★★ | ★★★★ | ★★★★★ | |
| GPT–3.5 / 4 | ★ | 24 / **38** | 0 / 0 | 0 / 0 | 0 / 0 | 0 / 0 | 100.0 / 100.0 |
| | ★★ | 71 / 59 | 60 / **70** | 16 / 24 | 0 / 4 | 0 / 0 | 40.8 / **44.6** |
| | ★★★ | 5 / 3 | 39 / 28 | **66** / 63 | 23 / 26 | 0 / 1 | 49.6 / **52.1** |
| | ★★★★ | 0 / 0 | 1 / 2 | 18 / 13 | **73** / 67 | 30 / 31 | 59.8 / 59.3 |
| | ★★★★★ | 0 / 0 | 0 / 0 | 0 / 0 | 4 / 3 | **70** / 68 | 94.6 / **95.8** |
| Precision | | 24.0 / 38.0 | 60.0 / 70.0 | 66.0 / 63.0 | 73.0 / 67.0 | 70.0 / 68.0 | % |

*Table 4.* Rating Performance Comparison for English language

| GPT–3.5 / 4 | Precision [%] | Recall [%] | F1 [%] |
|---|---|---|---|
| ★ | 24.0 / **38.0** | **100.0** / **100.0** | 38.7 / **55.1** |
| ★★ | 60.0 / **70.0** | 40.8 / **44.6** | 48.6 / **54.5** |
| ★★★ | **66.0** / 63.0 | 49.6 / **52.1** | 56.7 / **57.0** |
| ★★★★ | **73.0** / 67.0 | **59.8** / 59.3 | **65.8** / 62.9 |
| ★★★★★ | **70.0** / 68.0 | 94.6 / **95.8** | **80.5** / 79.5 |

*Table 5.* Overall Comparison for English language

| Version | Overall Accuracy [%] | Overall Error [%] | Cohen's kappa (k) | Correctly Classified | Incorrectly Classified | macro F1 [%] |
|---|---|---|---|---|---|---|
| GPT–3.5/4 | 58.6 / **61.2** | 41.4 / **38.8** | 0.49 / **0.52** | 293 / **306** | 207 / **194** | 58.1 / **61.8** |

*Table 6.* Evaluation German language GPT–3.5 and GPT–4

| | GPT–3.5 ⌀ | GPT–4 ⌀ | GPT–3.5 Δ | GPT–4 Δ | GPT–3.5 s | GPT–4 s | GPT–3.5 CSD | GPT–4 CSD |
|---|---|---|---|---|---|---|---|---|
| ★ | 2.16 | **1.74** | 1.16 | **0.74** | 0.526 | **0.505** | 288 | 213 |
| ★★ | 2.65 | 2.32 | 0.67 | **0.40** | 0.570 | **0.532** | | |
| ★★★ | **3.07** | 2.84 | **0.31** | 0.42 | **0.465** | 0.516 | | |
| ★★★★ | **3.90** | 3.83 | **0.30** | 0.35 | **0.503** | 0.539 | | |
| ★★★★★ | 4.56 | **4.78** | 0.44 | **0.22** | 0.519 | **0.484** | | |
| Mean | – | – | 0.58 | **0.43** | **0.517** | 0.515 | | |

*Table 7.* Confusion Matrix German Language GPT–3.5/4

| | | Rating | | | | | Recall |
|---|---|---|---|---|---|---|---|
| | | ★ | ★★ | ★★★ | ★★★★ | ★★★★★ | |
| GPT–3,5 / 4 | ★ | 7 / **29** | 1 / 4 | 0 / 1 | 0 / 0 | 0 / 0 | 87.5 / 85.3 |
| | ★★ | 70 / 68 | 38 / **62** | 12 / 27 | 2 / 3 | 0 / 1 | 31.2 / 38.5 |
| | ★★★ | 23 / 3 | 56 / 32 | **69** / 59 | 16 / 20 | 1 / 0 | 41.8 / 51.8 |
| | ★★★★ | 0 / 0 | 5 / 2 | 19 / 13 | **72** / 68 | 42 / 19 | 52.2 / 66.7 |
| | ★★★★★ | 0 / 0 | 0 / 0 | 0 / 0 | 10 / 9 | 57 / **80** | 85.1 / 89.9 |
| Precision | | 7.0 / 29.0 | 38.0 / 62.0 | 69.0 / 59.0 | 72.0 / 68.0 | 57.0 / 80.0 | % |

*Table 8.* Rating Performance Comparison for German language

| GPT–3.5 / 4 | Precision [%] | Recall [%] | F1 [%] |
|---|---|---|---|
| ★ | 7.0 / **29.0** | **87.5** / 85.3 | 13.0 / **43.3** |
| ★★ | 38.0 / **62.0** | 31.2 / **38.5** | 34.2 / **47.5** |
| ★★★ | **69.0** / 59.0 | 41.8 / **51.8** | 52.1 / **55.1** |
| ★★★★ | **72.0** / 68.0 | 52.2 / **66.7** | 60.5 / **67.3** |
| ★★★★★ | 57.0 / **80.0** | 85.1 / **89.9** | 68.3 / **84.7** |

*Table 9.* Overall Comparison for German language

| Version | Overall Accuracy [%] | Overall Error [%] | Cohen's kappa (k) | Correctly Classified | Incorrectly Classified | macro F1 [%] |
|---|---|---|---|---|---|---|
| GPT–3.5/4 | 48.6 / **59.6** | 51.4 / **40.4** | 0.36 / **0.50** | 243 / **298** | 257 / **202** | 45.6 / **59.6** |

one of the largest deviations of all languages and star categories with a deviation of 1.16 stars. However, GPT-4 increases the performance by around 30% and reduces the mean deviation to 0.74. With GPT-4, German has a very low average deviation of only 0.22 for 5-star ratings. In this category, 80/100 ratings were recognized correctly. The confusion matrix in Table 7 is influenced by the high deviation of GPT-3.5 described in Table 6.

This is primarily reflected by the values for precision. In this example, the difference within the precision is considerably higher than in other categories. The large deviation mentioned above also affects the results displayed in Table 8. The F1 value for the 1-star category decreases to 13% because of this particular outlier. However, by changing the GPT version, the overall accuracy can be increased by 11% (see Table 9). When being compared to English, the overall accuracy is slightly lower for German (see Table 8 and Table 9, respectively). Furthermore, the overall error for German is also a little greater than for English, which results also in the numbers of correct and incorrect classified reviews.

### *E. Results for Dutch*

Table 10 – Table 13 present the statistic evaluation (Table 10), the confusion matrix (Table 11), the rating performance comparison (Table 12) and the overall comparison of the two ChatGPT models for Dutch (Table 13).

According to the results in Table 10, 1-star ratings are consistently less well recognized by both versions when written in Dutch. The average deviation for GPT-3.5 and 4 already shows increased values compared to English and German, which already indicates an increased variance due to the language used.

No unexpected results can be described for the confusion matrix in Table 11. Merely in the 4-star category did GPT-3.5 slightly outperform the newer version GPT-4. Compared to English and German, lower F1 values in some star categories can be observed (Table 12).

According to the results in Table 13, in the case of Dutch there is a higher overall

*Table 10.* Evaluation Dutch GPT–3.5 and GPT–4

| | GPT–3.5 ⌀ | GPT–4 ⌀ | GPT–3.5 Δ | GPT–4 Δ | GPT–3.5 s | GPT–4 s | GPT–3.5 CSD | GPT–4 CSD |
|---|---|---|---|---|---|---|---|---|
| ★ | 1.89 | **1.58** | 0.89 | **0.58** | **0.567** | 0.638 | 296 | **227** |
| ★★ | 2.49 | **2.18** | 0.51 | **0.38** | **0.522** | 0.528 | | |
| ★★★ | 2.70 | **2.71** | 0.58 | **0.51** | **0.516** | 0.541 | | |
| ★★★★ | 3.70 | **3.80** | **0.48** | 0.52 | 0.611 | **0.559** | | |
| ★★★★★ | 4.50 | **4.72** | 0.50 | **0.28** | 0.503 | **0.451** | | |
| Mean | – | – | 0.59 | **0.45** | 0.544 | **0.543** | | |

*Table 11.* Confusion Matrix Dutch GPT–3.5 and GPT–4

| | | Rating | | | | | Recall |
|---|---|---|---|---|---|---|---|
| | | ★ | ★★ | ★★★ | ★★★★ | ★★★★★ | |
| GPT–3.5 / 4 | ★ | 22 / **50** | 1 / 10 | 1 / 2 | 0 / 0 | 0 / 0 | 91.7 / 80.7 |
| | ★★ | 67 / 42 | 50 / **64** | 42 / 36 | 6 / 3 | 0 / 0 | 30.3 / 44.1 |
| | ★★★ | 11 / 8 | 48 / 24 | 43 / **51** | 27 / 30 | 0 / 0 | 33.3 / 45.1 |
| | ★★★★ | 0 / 0 | 1 / 2 | 14 / 11 | **58** / 51 | 50 / 28 | 47.2 / 55.4 |
| | ★★★★★ | 0 / 0 | 0 / 0 | 0 / 0 | 9 / 16 | 50 / **72** | 84.8 / 81.8 |
| Precision | | 22.0 / 50.0 | 50.0 / 64.0 | 43.0 / 51.0 | 58.0 / 51.0 | 50.0 / 72.0 | % |

*Table 12.* Rating Performance Comparison for Dutch language

| GPT–3.5 / 4 | Precision [%] | Recall [%] | F1 [%] |
|---|---|---|---|
| ★ | 22.0 / **50.0** | **91.7** / 80.7 | 35.5 / **61.7** |
| ★★ | 50.0 / **64.0** | 30.3 / **44.1** | 37.7 / **52.2** |
| ★★★ | 43.0 / **51.0** | 33.3 / **45.1** | 37.6 / **47.9** |
| ★★★★ | **58.0** / 51.0 | 47.2 / **55.4** | 52.0 / **53.1** |
| ★★★★★ | 50.0 / **72.0** | **84.8** / 81.8 | 62.9 / **76.6** |

*Table 13.* Overall Comparison for Dutch language

| Version | Overall Accuracy [%] | Overall Error [%] | Cohen's kappa (k) | Correctly Classified | Incorrectly Classified | macro F1 [%] |
|---|---|---|---|---|---|---|
| GPT–3.5/4 | 44.6 / **57.6** | 55.4 / **42.4** | 0.31 / **0.47** | 223 / **288** | 277 / **212** | 45.1 / **58.3** |

accuracy for GPT-4 with 57.6% compared to GPT-3.5 with 44.6%. Equally, the overall error decreased with the updated GPT-4 version compared to GPT-3.5.

### *F. Results for Korean*

Table 14 – Table 17 give the statistic evaluation (Table 14), the confusion matrix (Table 15), the rating performance comparison (Table 16) and the overall comparison of the two ChatGPT models in the case of Korean (Table 17).

Here too, the average deviations are consistently higher than in Dutch, German and English. For version GPT-3.5, Table 14 shows again that 1-star ratings are rated lower than in the other categories. The performance with 5-star ratings of both versions in Korean is outstanding, where the smallest deviations with values of 0.589 and 0.584 could be detected. Moreover, Table 15 illustrates the small deviation with 79/100 and 80/100 correctly recognized 5-star ratings, respectively, which can be explained by the way reviews are written, especially in the Korean language and culture.

First, the advantages are emphasized relatively clearly and unambiguously in the body text and the disadvantages are explained less concisely. At the end of many of these reviews, a categorical position is often taken on a product's features. According to the results in Table 16, the 7.2% increase in accuracy with

*Table 14.* Evaluation Korean language GPT–3.5 and GPT–4

| | GPT–3.5 ⌀ | GPT–4 ⌀ | GPT–3.5 Δ | GPT–4 Δ | GPT–3.5 s | GPT–4 s | GPT–3.5 CSD | GPT–4 CSD |
|---|---|---|---|---|---|---|---|---|
| ★ | 1.92 | **1.58** | 0.92 | **0.58** | 0.800 | **0.727** | 310 | 263 |
| ★★ | 2.48 | **2.21** | 0.62 | **0.53** | 0.632 | **0.611** | | |
| ★★★ | **2.89** | 2.68 | 0.79 | **0.76** | **0.478** | 0.571 | | |
| ★★★★ | **3.88** | 3.84 | **0.56** | **0.56** | 0.625 | **0.608** | | |
| ★★★★★ | 4.79 | **4.80** | 0.21 | **0.20** | 0.409 | **0.402** | | |
| Mean | – | – | 0.62 | **0.53** | 0.589 | **0.584** | | |

*Table 15.* Confusion Matrix Korean language GPT–3.5/4

| | | Rating | | | | | Recall |
|---|---|---|---|---|---|---|---|
| | | ★ | ★★ | ★★★ | ★★★★ | ★★★★★ | |
| GPT–3,5 / 4 | ★ | 28 / **50** | 7 / 16 | 1 / 5 | 0 / 0 | 0 / 0 | 77.8 / 70.4 |
| | ★★ | 58 / 46 | 46 / **53** | 43 / 44 | 7 / 6 | 0 / 0 | 29.9 / 35.6 |
| | ★★★ | 10 / 2 | 39 / 25 | 24 / **31** | 20 / 24 | 0 / 0 | 25.8 / 37.8 |
| | ★★★★ | 2 / 0 | 8 / 6 | 30 / 18 | **51** / 50 | 21 / 20 | 45.5 / 53.2 |
| | ★★★★★ | 2 / 2 | 0 / 0 | 2 / 2 | 22 / 20 | 79 / **80** | 75.2 / 76.9 |
| Precision | | 28.0 / 50.0 | 46.0 / 53.0 | 24.0 / 31.0 | 51.0 / 50.0 | 79.0 / 80.0 | % |

*Table 16.* Rating Performance Comparison for Korean language

| GPT–3.5 / 4 | Precision [%] | Recall [%] | F1 [%] |
|---|---|---|---|
| ★ | 28.0 / **50.0** | **77.8** / 70.4 | 41.2 / **58.5** |
| ★★ | 46.0 / **53.0** | 29.9 / **35.6** | 36.2 / **42.6** |
| ★★★ | 24.0 / **31.0** | 25.8 / **37.8** | 24.9 / **34.1** |
| ★★★★ | **51.0** / 50.0 | 45.5 / **53.2** | 48.1 / **51.6** |
| ★★★★★ | 79.0 / **80.0** | 75.2 / **76.9** | 77.1 / **78.4** |

*Table 17.* Overall Comparison for Korean language

| Version | Overall Accuracy [%] | Overall Error [%] | Cohen's kappa (k) | Correctly Classified | Incorrectly Classified | macro F1 [%] |
|---|---|---|---|---|---|---|
| GPT–3.5/4 | 45.6 / **52.8** | 54.4 / **47.2** | 0.32 / **0.41** | 228 / **264** | 272 / **236** | / **53.0** |

GPT-4 is already reduced compared to Dutch (13%).

According to the results in Table 17, the overall accuracy improves for GPT-4 compared to GPT-3.5. Furthermore, the overall error again decreases by 7.2%.

### *G. Results for Hindi*

For Hindi, Table 18 – Table 21 show the statistic evaluation (Table 18), the confusion matrix (Table 19), the rating performance comparison (Table 20) and the overall comparison of the two ChatGPT models (Table 21). In this language, the values shown in Table 18 have the highest mean deviation of all the languages examined.

The dispersion is also significantly higher here. These deviations can also be seen in the confusion matrix in Table 19. GPT-4 increased the overall accuracy to 47.8%, a slight increase of 4% compared to GPT-3.5 (see Table 21)

According to the results in Table 20, the precision and recall are significantly lower in Hindi compared to the other four languages examined. Hence, also the F1-values are slightly poorer than for Englisch, German, Durch and Korean.

However, when comparing GPT-3.5 to GPT-4 for Hindi the overall accuracy increases for GPT-4 while the overall error again decreases.

*Table 18.* Evaluation Hindi language GPT–3.5 and GPT–4

| | GPT–3.5 ⌀ | GPT–4 ⌀ | GPT–3.5 Δ | GPT–4 Δ | GPT–3.5 s | GPT–4 s | GPT–3.5 CSD | GPT–4 CSD |
|---|---|---|---|---|---|---|---|---|
| ★ | 1.87 | **1.56** | 0.87 | **0.56** | 0.861 | **0.770** | 342 | 304 |
| ★★ | 2.22 | **1.98** | **0.44** | 0.48 | 0.641 | **0.559** | | |
| ★★★ | **2.99** | **3.01** | **0.85** | **0.85** | **0.642** | 0.657 | | |
| ★★★★ | **4.07** | 4.19 | **0.63** | 0.65 | 0.646 | **0.520** | | |
| ★★★★★ | 4.37 | **4.50** | 0.63 | **0.50** | 0.825 | **0.772** | | |
| Mean | – | – | 0.68 | **0.61** | 0.723 | **0.656** | | |

*Table 19.* Confusion Matrix Hindi language GPT–3.5 and GPT–4

| | | Rating | | | | | Recall |
|---|---|---|---|---|---|---|---|
| | | ★ | ★★ | ★★★ | ★★★★ | ★★★★★ | |
| GPT–3,5 / 4 | ★ | 31 / **55** | 11 / 25 | 4 / 5 | 2 / 0 | 1 / 1 | 63.3 / 64.0 |
| | ★★ | 60 / 38 | **62** / 55 | 35 / 32 | 3 / 2 | 3 / 2 | 38.0 / 42.6 |
| | ★★★ | 4 / 5 | 23 / 17 | 29 / **30** | 16 / 19 | 7 / 5 | 36.7 / 39.5 |
| | ★★★★ | 1 / 0 | 2 / 3 | 22 / 23 | **44** / 37 | 36 / 30 | 41.9 / 39.8 |
| | ★★★★★ | 4 / 2 | 2 / 0 | 10 / 10 | 35 / 42 | 53 / **62** | 51.0 / 53.5 |
| Precision | | 31.0 / 55.0 | 62.0 / 55.0 | 29.0 / 30.0 | 44.0 / 37.0 | 53.0 / 62.0 | % |

*Table 20.* Rating Performance Comparison for Hindi language

| GPT–3.5 / 4 | Precision [%] | Recall [%] | F1 [%] |
|---|---|---|---|
| ★ | 31.0 / **55.0** | 63.3 / **64.0** | 41.6 / **59.1** |
| ★★ | **62.0** / 55.0 | 38.0 / **42.6** | 47.2 / **48.0** |
| ★★★ | 29.0 / **30.0** | 36.7 / **39.5** | 32.4 / **34.1** |
| ★★★★ | **44.0** / 37.0 | **41.9** / 39.8 | **42.9** / 38.3 |
| ★★★★★ | 53.0 / **62.0** | 51.0 / **53.5** | 52.0 / **57.4** |

*Table 21.* Overall Comparison for Hindi language

| Version | Overall Accuracy [%] | Overall Error [%] | Cohen's kappa (k) | Correctly Classified | Incorrectly Classified | macro F1 [%] |
|---|---|---|---|---|---|---|
| GPT–3.5/4 | 43.8 / **47.8** | 56.2 / **52.2** | 0.30 / **0.35** | 219 / **239** | 281 / **261** | 43.2 / **47.4** |

## 5. EVALUATION OF THE RESULTS

According to Statista 52.1% of the websites use English, while German is used on 4.8% of the websites, 1.8% of the websites are Dutch, 0.8% are Korean and the number of websites in Hindi is even smaller (Statista 2024). Therefore, English is considered and treated as a reference group. German and Dutch represent websites over 1% on the internet. German is at the upper end and Dutch at the lower end of this spectrum. Therefore, the mean values of deviation and standard deviation for each language have been calculated and are listed in the results and findings section. These two languages (German and Dutch) will form the represented language group (REP) with and denoting the deviation and the standard deviation respecively. For Korean and Hindi, which are underrepresented languages (proportion < 1%), the same method is applied, and they form the unrepresented language group (UREP) with and for the deviation and the standard deviation respectively.

### *A. Language comparison*

For the statistical comparison of the language groups, a Levene test was applied to

test the homogeneity of variance. The first Levene test was conducted among English, REP and UREP. Since the variances between all groups proved to be different as $p_{L1} = 4.39005 \cdot 10^{-7}$ compared to the chosen significance level of $\alpha = 0.05$,, no ordinary two-sample t-tests could be carried out. Instead, two independent-sample t-tests were carried out under the assumption of unequal variances (Ruxton 2006).

The first t-test compared English (en) to represented language group (REP) and the second one compared the represented language group (REP) to the unrepresented language group (UREP). The basic hypothesis is that the more represented a language group is on the internet, the smaller is the deviation ($\Delta_{en} < \Delta_{REP} < \Delta_{UREP}$). In the case of multiple tests, however, alpha error accumulation must be considered. For two tests $\alpha = 1 - (1 - 0.05)^2 = 0.0975$ is used. To counteract this, the Bonferroni correction is applied (Benjamini and Hochberg 1995), which states that the significance level needs to be divided by the number of tests performed in order to reduce the individual probability of a single hypothesis test. Table 22 shows that both p-values are below the corrected significance level.

Consequently, the previously formulated basic hypothesis, which states that the deviation in the evaluation by GPT is smaller provided the more represented a language or language group is on the internet (hypothesis A1), must be accepted.

### *B. Version comparison*

The F1 values determined are shown in Fig. 5. The top bar indicates the corresponding distribution of a language on the internet in percentage. The middle bar depicts the macro F1 value achieved for GPT-3.5 and the one below for GPT-4.

*Table 22.* Two–sample t–test unequal variances en–REP/REP–UREP GPT–3.5

| **Left–sided two–sample t–test en–REP for unequal variances** | **Left–sided two–sample t–test REP–UREP for unequal variances** |
|---|---|
| Alternative hypothesis: $\Delta_{en} - \Delta_{REP} < 0$ | Alternative hypothesis: $\Delta_{REP} - \Delta_{UREP} < 0$ |
| $\alpha = 0.025$ | $\alpha = 0.025$ |
| $p_{T1} = 6.21 \cdot 10^{-8}$ | $p_{T2} = 9.20 \cdot 10^{-3}$ |

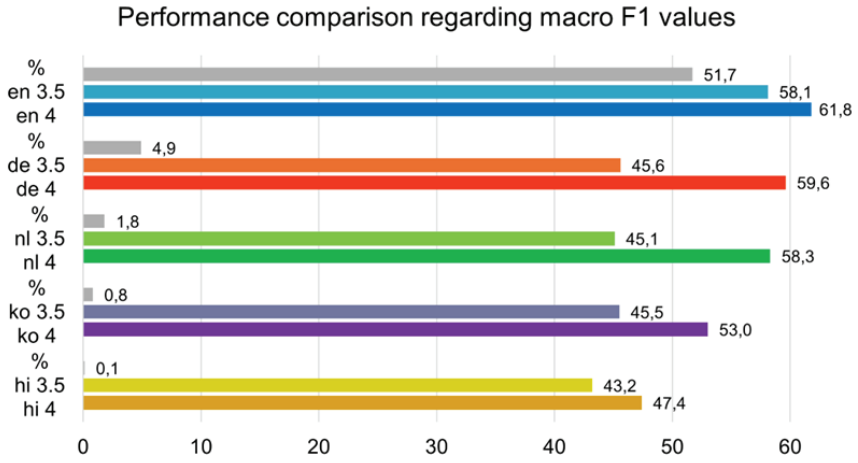


*Figure 5.* Performance comparison macro F1 value between GPT–3.5 and GPT–4

*Table 23.* Cumulative star deviation of all languages GPT–3.5 and GPT–4

| CSD | **GPT–3.5 Σ (in stars)** | **GPT–4 Σ (in stars)** | **Δ GPT–3.5 to GPT–4 (in %)** |
|---|---|---|---|
| en | 213 | 208 | 2.35 |
| de | 288 | 213 | 26.04 |
| nl | 296 | 227 | 30.39 |
| ko | 342 | 263 | 17.87 |
| hi | 342 | 304 | 12.5 |

These values diverge more clearly when switching to the more powerful model. If all languages are considered, the more powerful model shows consistently higher macro F1 values. This is also confirmed by the CSD parameter between GPT-3.5 and GPT-4 depending on the language used depicted in Table 23.

When comparing the CSD values to the presented F1 values in Fig. 5 there is a similar decline in accuracy depending on the language used.

Prior to proving a significant difference between GPT-3.5 and GPT-4 preliminary Shapiro-Wilk-Tests to check for normality must be conducted (Table 24).

Since both $p_{S1}$ and $p_{S2}$ in Table 24 are not significant, it can be assumed that the F1 and CSD values for both versions are normally distributed. Since normal distribution is present and both models evaluate the same sample. i.e. all reviews, paired t-tests are used for version comparison (Table 25). Both paired t-tests in Table 25 show a significant difference for F1 and CSD values upon both versions. This proves a superior performance of GPT-4 over GPT-3.5 thus supporting hypothesis A2.

### *C. Context length*

Only the results of GPT-4 are used for the statistical comparison regarding context length. The following differences exist between REP and UREP regarding the average length of reviews.

The first assumption examined was that ratings in German have on average double the number of words compared to Dutch. The second assumption was that ratings in Korean have on average three times the word count of Hindi.

First, Levene tests are used again to check the homogeneity of variances between the languages within REP (German to Dutch) and UREP (Korean to Hindi). The results of both Levene tests, therefore, determine which two-sample t-test approach is used. Table 26 shows no different variances between German and Dutch ($p_{L2} = 0.196$) and between Korean and Hindi ($p_{L3} = 0.186$), so ordinary

*Table 24.* Shapiro–Wilk test F1/CSD

| **Shapiro Wilk F1** | **Shapiro Wilk CSD** |
|---|---|
| Alternative hypothesis: $F1_{GPT-3.5}$, $F1_{GPT-4} \neq N(\mu, \sigma^2)$ | Alternative hypothesis: $CSD_{GPT-3.5}$, $CSD_{GPT-4} \neq N(\mu, \sigma^2)$ |
| $\alpha = 0.05$ | $\alpha = 0.05$ |
| $W = 0.857$ | $W = 0.854$ |
| $p_{S1} = 0.219$ | $p_{S2} = 0.209$ |

*Table 25.* Paired t–test F1/CSD

| **Paired t–test F1** | **Paired t–test CSD** |
|---|---|
| Alternative hypothesis: $\Delta_{GPT-3.5,\ GPT-4} \neq 0$ | Alternative hypothesis: $\Delta_{GPT-3.5,\ GPT-4} \neq 0$ |
| $\alpha = 0.05$ | $\alpha = 0.05$ |
| $p_{T3} = 0.008$ | $p_{T4} = 0.009$ |

*Table 26.* Levene test de–nl/ko–hi GPT–4

| **Levene test de–nl for homogeneity of variance** | **Levene test ko–hi for homogeneity of variance** |
|---|---|
| Alternative hypothesis: $s^2_{de} \neq s^2_{nl}$ | Alternative hypothesis: $s^2_{ko} \neq s^2_{hi}$ |
| $\alpha = 0.05$ | $\alpha = 0.05$ |
| $p_{L2} = 0.196$ | $p_{L3} = 0.186$ |

*Table 27.* Two–sample t–test equal variances de–nl/ko–hi GPT–4

| **Left–sided two–sample t–test de–nl for equal variances** | **Left–sided two–sample t–test ko–hi for equal variances** |
|---|---|
| Alternative hypothesis: $\Delta_{de} - \Delta_{nl} < 0$ | Alternative hypothesis: $\Delta_{ko} - \Delta_{hi} < 0$ |
| $\alpha = 0.05$ | $\alpha = 0.05$ |
| $p_{T5} = 0.210$ | $p_{T6} = 0.023$ |

*Table 28.* Hypotheses A1–A3: Review

| **Hypothesis A1** | | |
|---|---|---|
| Decision | Test | p–values (α = 0.025) |
| Accepted | Two–sample t–test unequal variances | $6.21 \cdot 10^{-8}$ and $9.20 \cdot 10^{-3}$ |
| **Hypothesis A2** | | |
| Decision | Test | p–values (α = 0.05) |
| Accepted | paired t-test F1/CSD | 0.008 and 0.009 |
| **Hypothesis A3** | | |
| Decision | Test | p–values (α = 0.05) |
| Discarded | Two–sample t–test equal variances de–nl/ko–hi | 0.210 and 0.023 |

independent-sample t-tests can, therefore, be used under the assumption of equal variances.

The significance level does not need to be adjusted here, as the left-sided two-sample tests for German/Dutch and Korean/Hindi are not linked and are, therefore, independent of each other. (Table 27).

Since only the left-sided two sample t-test for Korean/Hindi in Table 27 shows a significant p-value, the hypothesis that a higher word count in a review in unpresented languages leads to increased accuracy of GPT-4 must, therefore, be rejected, which relates to hypothesis A3. Consequently, hypothesis A3 can be partly supported.

### *D. Verification of the hypotheses*

After testing all the defined hypotheses, the results are listed in tabular form in Table 28.

Both p-values for hypothesis A1 are below the corrected significance level. Hence, for hypothesis A1, the claim can be made that the more frequently a language or a language within the three defined language groups is used on the internet, the more accurate the evaluation of e-commerce reviews by Chat GPT-3.5 is. In this study this statement is made under the relation of $\Delta_{EN} < \Delta_{REP} < \Delta_{UREP}$.

All macro F1 values consistently increased in all languages when GPT-4 is compared to GPT-3.5, although GPT-4 showed a lower total overall CSD value (1215 stars deviation) compared to GPT-3.5 (1149 stars deviation). Since the results for hypothesis A2 showed a significant difference upon these values it can be assumed that GPT-4 achieves a higher performance in evaluation reviews in different languages.

Hypothesis A3 is partly supported or discarded because the p-value is only statistically significant Korean/Hindi and not for German/Dutch. It can, therefore, be concluded that the context length in this study has no influence on the accuracy of GPT-4's evaluation of reviews in less represented languages.

## 6. CONCLUSION

### *A. Comparison between the languages examined*

The evaluation shows significantly that GPT-3.5 and GPT-4 recognize product reviews with varying degrees of accuracy in the languages examined. The selected language in which a review was written influences how strong the deviation in the evaluation by ChatGPT is. Fig. 6 shows the deviations (zero to four stars) for all 500 reviews in the respective languages and versions. For GPT-3.5, 98.8% and 91.4% of the deviation in English and in Hindi, respectively, is a maximum of one star. Languages such as Korean and Hindi show stronger tendencies with higher deviations (orange, dark red).

The results reflect the fact that the training data is the most important factor for accuracy. The training data of GPT-3.5 mainly comes from data from the internet, which was written in a few languages and thus influences the results.

There is no information on the training data mix for GPT-4, but a similar language distribution seems obvious. The greater complexity of the writing systems in Korean and Hindi could also have had an influence on the present results.

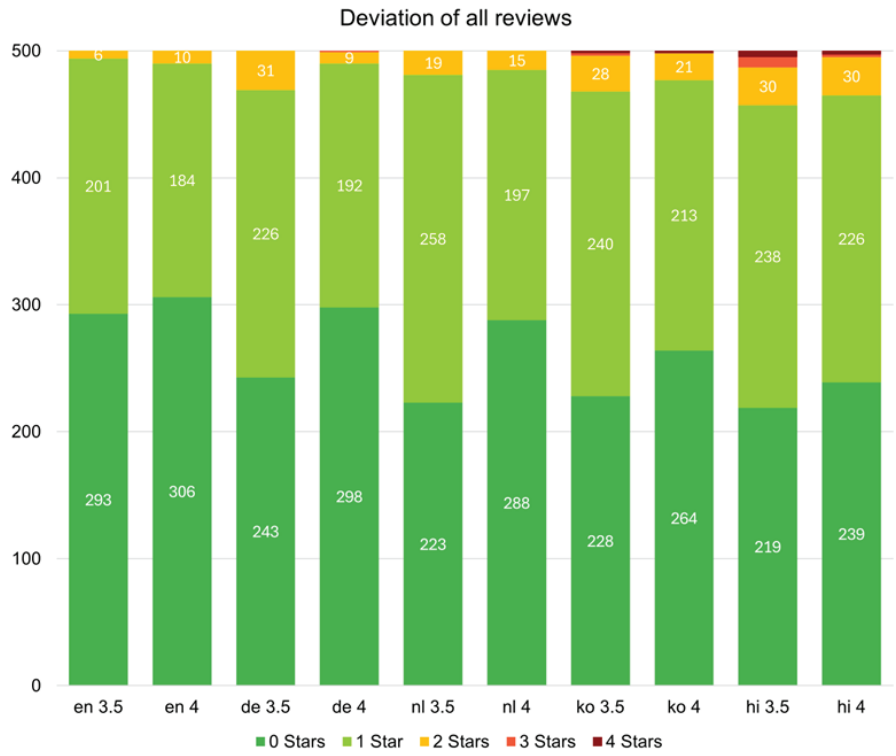


*Figure 6.* Deviations of all reviews for GPT–3.5 and GPT–4

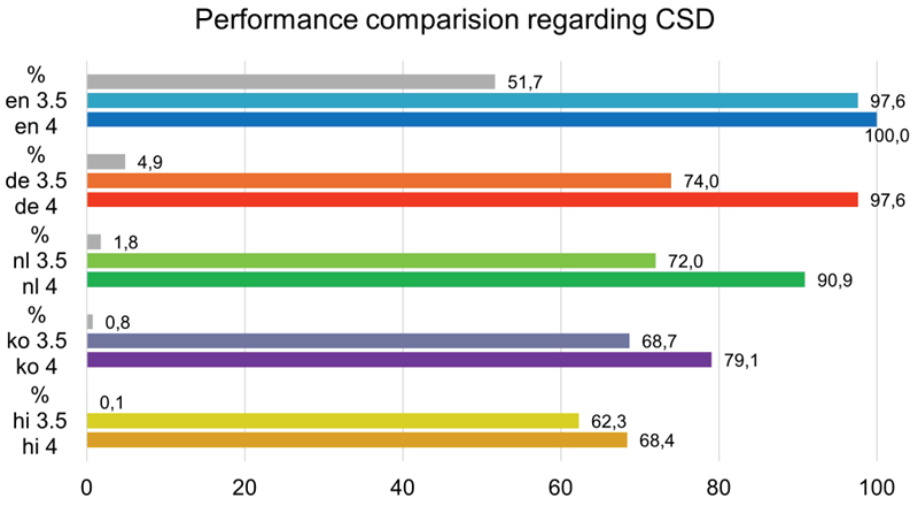


*Figure 7.* Performance comparison with CSD GPT–3.5 and GPT–4

### *B. Comparison between ChatGPT-3.5 and ChatGPT-4*

The macro F1 value as well as the cumulative star deviation show that GPT-4 evaluates and classifies more precisely. The best performance (CSD for English with GPT-4 = 208) is considered as a reference and benchmark and is set in relation to the other results, which results in a stronger differentiation as depicted in Fig. 7.

The highest increase in accuracy was recorded in German and Dutch when the model was changed. In addition, the performance of both models decreases continuously with the distribution level of a language on the internet.

### *C. Relationship between context length and effectiveness*

There is no statistical correlation between context length and effectiveness, which means, that more words in a review, in a direct comparison between German and Dutch and between Korean and Hindi, do not lead to a higher accuracy in the evaluation by GPT-4. If the results are visualized, further findings emerge. Fig 8. **Fehler! Verweisquelle konnte nicht**

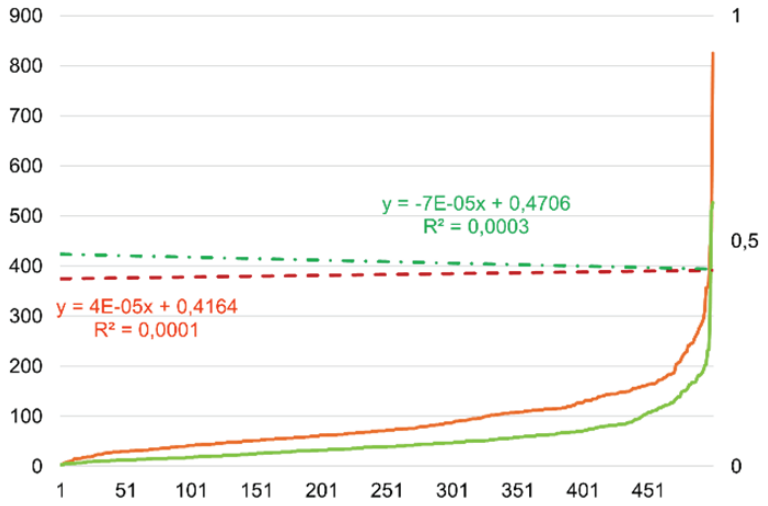


*Figure 8.* Trend lines of the deviations between de–nl GPT–4

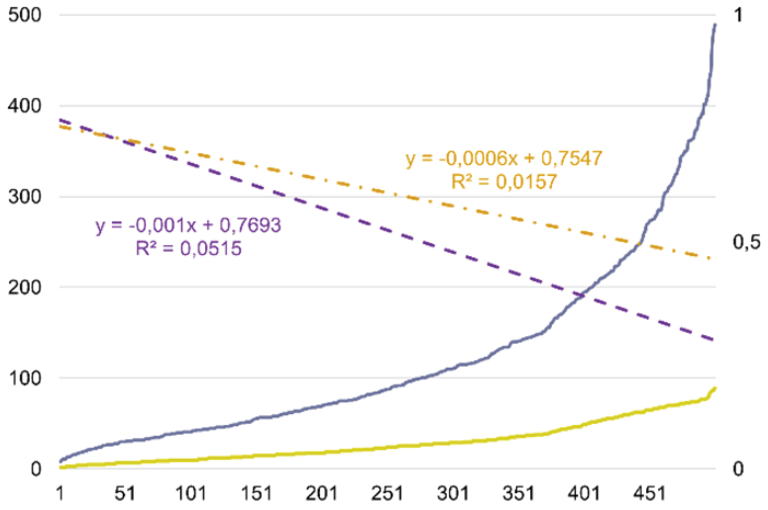


*Figure 9.* Trend lines of the deviations between ko–hi GPT–4

**gefunden werden.** and Fig. 9 depict trend lines of GPT-4's deviation in its evaluation in accordance with the word count across all four languages.

The regression coefficients in Fig. 8 and Fig. 9 are in the range from $10^{-5}$ to $10^{-3}$ and thus have a very small influence. Also, both R2 values in Fig. 8. and Fig. 9, i.e. the coefficients of determination of all regression lines, are close to 0, which indicates that the independent variable (word count of a review) has no influence on the dependent variable (deviation in the evaluation by GPT-4).

Therefore, these values show that a linear regression model cannot adequately determine the variability of this independent variable, which further strengthens the results of the statistical analysis to discard hypothesis A3. Since a $\beta = 0.2$ was chosen, there is a probability of 20% that this decision was incorrect.

### *D. Answering the research questions*

In this subsection the research question and subquestions are answered.

**RsubQ1:** Are more represented languages more accurately understood or recognized by ChatGPT-3.5 than less represented languages?

Yes, reviews written in more represented languages such as English are recognized significantly more accurately. The accuracy correlates with the prevalence of the language on the internet; German and Dutch perform better than Korean and Hindi.

**RSubQ2**: Which of the two models examined, ChatGPT-3.5 or 4, is more accurate in recognizing the sentiment of different natural languages?

ChatGPT-4 is more accurate than ChatGPT-3.5. The results show that GPT-4 recognizes semantic differences more accurately and achieves respectable results in

unrepresented languages, despite lower proportions of these languages within the training data.

**RSubQ3**: For less represented languages, does more context result in a more accurate assessment with ChatGPT-4?

No, more words do not result in higher accuracy. The number of words, therefore, has no significant influence on the accuracy of the ChatGPT rating.

**RQ**: How does language choice influence the effectiveness of ChatGPT in recognizing sentiment for review texts in different languages?

The effectiveness of sentiment recognition and language comprehension depends significantly on the language used. Languages with a high prevalence, such as English, are recognized more precisely by ChatGPT than languages with a low prevalence, such as Korean and Hindi, which show higher deviations and lower macro F1 values. Possible inaccurate tokenization for more complex writing systems in Korean and Hindi may also affect performance. ChatGPT-4 consistently achieves higher performance across all languages compared to ChatGPT-3.5. Simply increasing the word count in reviews does not necessarily lead to more accurate scoring in less represented languages.

### *D. Critical consideration and limitations*

The manual collection of the 2,500 reviews in this paper shows a pseudorandomized approach but could be made more efficient by web crawlers. A bias in the selection of German and English reviews is possible, as these languages are spoken and understood by the author. By using OpenAI-API and web crawlers, a higher reproducibility of this work could be achieved. Furthermore, the performance in a single NLU task should not serve as the sole criterion for language comprehension, but the results presented underline the role of the language used, which allows conclusions to be drawn about the multilingual language comprehension of ChatGPT.

### *E. Conclusion and outlook*

The questions posed were answered satisfactorily. Future work could benefit from the aforementioned more efficient methodology and comparisons with other models such as Gemini or LLaMa. A direct comparison of the classification with human test subjects can also be informative. The extension to existing data sets from platforms such as Trustpilot, Expedia or social media could provide further insights. With social media, however, attention must be paid to data quality. As performance is declining in less common languages, adapted models are already being used to counteract this. In China, for example, QWEN-1.5 (Qian Wen) is used by Alibaba as a model with an adapted training data mix for specific languages such as Chinese (Bai et al. 2023; Github 2024a). Chinese language-specific models have also been developed (Zhu and Luo 2022).

## 7. SUMMARY

In this study, the multilingual language comprehension of ChatGPT-3.5 and 4 was examined. This study was carried out under the premise that web pages were only written in a few languages – which in turn represent part of ChatGPT's training data. Therefore, their performance was compared with a sentiment analysis depending on the frequency of a language on the internet. For this purpose, 500 e-commerce reviews were collected in each of the languages English, German, Dutch, Korean and Hindi. English is dominant on the internet with over 50% representation, while German and Dutch have more than 1% and Korean and Hindi less than 1% representation. These 500 ratings in one language are evenly divided into 100 one to five star ratings each. All ratings were shown to both versions of ChatGPT, which were then asked to rate and categorize them within one to five stars. The outputs were then analyzed using statistical methods and metrics based on two-sample t-tests, confusion matrices, macro F1 values and the sum of the star deviations for all languages and also presented graphically.

The results show a statistically significant correlation between the selected language and the accuracy of the evaluation of reviews by ChatGPT-3.5. It turns out that the more represented a language of this study is on the internet, the more effective the classification is. A performance comparison between version 3.5 and version 4 shows that ChatGPT-4 consistently achieves higher macro F1 values and a lower cumulative star deviation. This indicates a more precise classification in these languages. The deviations of the ratings by ChatGPT-4 for German and Dutch are comparable to those in English. Finally, the potential influence of longer reviews in German, Dutch, Korean and Hindi, which are used to varying

degrees on the internet, was investigated. It was found that an increase in the number of words in these lesser represented languages does not lead to a higher accuracy in the classification by ChatGPT-4. The results obtained, therefore, suggest that the performance of the underlying large language models, as with GPT, may show significant differences in sentiment analysis in a multilingual context.

This study shows the importance of multilingualism in the future development and optimization of large language models. With the recently released GPT-4o, OpenAI promises support for more languages than before. Full multimodality is now also available with ChatGPT-4o (use of audio, images, files, videos and camera on both desktop and smartphone). GPT-4o is available free of charge, so it can be assumed that there will be an increase in users who also want to use ChatGPT in their native language. This underlines the importance of multilingualism, as these new users sometimes use languages that are not fully understood or supported. Future research and work should also consider approaches such as expanding training datasets and applying transfer learning. This not only increases accuracy, but also fairness towards other languages. Ultimately, companies can also benefit from the further development of these models if factors such as inclusive use or linguistic and cultural differences are taken into account.

## REFERENCES


Achiam, Josh; Adler, Steven; Agarwal, Sandhini; Ahmad, Lama; Akkaya, Ilge; Aleman, Florencia Leoni et al. (2023): GPT-4 Technical Report. Available online at http://arxiv.org/pdf/2303.08774.

Ahuja, Kabir; Diddee, Harshita; Hada, Rishav; Ochieng, Millicent; Ramesh, Krithika; Jain, Prachi et al. (2023): MEGA: Multilingual Evaluation of Generative AI. Available online at http://arxiv.org/pdf/2303.12528.

Alec Radford; Jeff Wu; R. Child; D. Luan; Dario Amodei; I. Sutskever (2019): Language Models are Unsupervised Multitask Learners. Available online at https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe.

Alec Radford; Karthik Narasimhan (2018): Improving Language Understanding by Generative Pre-Training. Available online at https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035.

Bai, Jinze; Bai, Shuai; Chu, Yunfei; Cui, Zeyu; Dang, Kai; Deng, Xiaodong et al. (2023): Qwen Technical Report. Available online at http://arxiv.org/pdf/2309.16609.

Benjamini, Yoav; Hochberg, Yosef (1995): Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. In *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1), pp. 289–300. Available online at http://www.jstor.org/stable/2346101.

Bishop, Christopher M.; Bishop, Hugh (2024): Deep learning. Foundations and concepts. Cham: Springer. Available online at https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=30853138.

Brown, Tom B.; Mann, Benjamin; Ryder, Nick; Subbiah, Melanie; Kaplan, Jared; Dhariwal, Prafulla et al. (2020): Language Models are Few-Shot Learners. Available online at http://arxiv.org/pdf/2005.14165.

Brundage, Miles; Avin, Shahar; Wang, Jasmine; Belfield, Haydn; Krueger, Gretchen; Hadfield, Gillian et al. (2020): Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. Available online at http://arxiv.org/pdf/2004.07213.

Cekuls, Andrejs (2023): AI-Driven Competitive Intelligence: Enhancing Business Strategy and Decision Making. In *Journal of Intelligence Studies in Business*, 12 (3), pp. 4–5. https://doi.org/10.37380/jisib.v12i3.961.

Das, Mithun; Pandey, Saurabh Kumar; Mukherjee, Animesh (2023): Evaluating ChatGPT's Performance for Multilingual and Emoji-based Hate Speech Detection. Available online at http://arxiv.org/pdf/2305.13276.

Das, Rupak Kumar; Pedersen, Ted (2024): SemEval-2017 Task 4: Sentiment Analysis in Twitter using BERT. Available online at http://arxiv.org/pdf/2401.07944.

Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina (2018): BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Available online at http://arxiv.org/pdf/1810.04805.

Döring, Nicola (2023): Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. 6., vollständig überarbeitete, aktualisierte und erweiterte Auflage. Berlin: Springer (Springer-Lehrbuch).

Erös, Bernhard (2024): Reviews and evaluation of ChatGPT. Available online at https://www.researchgate.net/search/researcher?q=bernhard%

2Ber%25C3%25B6s, updated on 6/23/2024, checked on 6/23/2024.

Github (2024a): GitHub – QwenLM/Qwen2: Qwen2 is the large language model series developed by Qwen team, Alibaba Cloud. Available online at https://github.com/QwenLM/Qwen2, updated on 6/23/2024, checked on 6/23/2024.

Github (2024b): Google Research Multilingual Bidirectional Encoder Representations from Transformers (BERT). Available online at https://github.com/google-research/bert/blob/master/multilingual.md#list-of-languages, updated on 6/23/2024, checked on 6/23/2024.

Glaese, Amelia; McAleese, Nat; Trębacz, Maja; Aslanides, John; Firoiu, Vlad; Ewalds, Timo et al. (2022): Improving alignment of dialogue agents via targeted human judgements. Available online at http://arxiv.org/pdf/2209.14375.

Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome H. (2017): The elements of statistical learning. Data mining, inference, and prediction. Second edition. New York, NY: Springer (Springer Series in Statistics). Available online at https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=6314834.

Hello GPT-4o (2024). Available online at https://openai.com/index/hello-gpt-4o/, updated on 6/23/2024, checked on 6/23/2024.

Hendrycks, Dan; Burns, Collin; Basart, Steven; Critch, Andrew; Li, Jerry; Song, Dawn; Steinhardt, Jacob (2020a): Aligning AI With Shared Human Values. Available online at http://arxiv.org/pdf/2008.02275.

Hendrycks, Dan; Burns, Collin; Basart, Steven; Zou, Andy; Mazeika, Mantas; Song, Dawn; Steinhardt, Jacob (2020b): Measuring Massive Multitask Language Understanding. Available online at http://arxiv.org/pdf/2009.03300.

Henighan, Tom; Kaplan, Jared; Katz, Mor; Chen, Mark; Hesse, Christopher; Jackson, Jacob et al. (2020): Scaling Laws for Autoregressive Generative Modeling. Available online at http://arxiv.org/pdf/2010.14701.

Hung Vo, Trung; Felde, Imre; Ninh, Khanh Chi (2025): Fake News Detection System, based on CBOW and BERT. In *ACTA POLYTECH HUNG* 22 (1), pp. 27–41. https://doi.org/10.12700/APH.22.1.2025.1.2.

Jin, Hongpeng; Wei, Wenqi; Wang, Xuyu; Zhang, Wenbin; Wu, Yanzhao (2023): Rethinking Learning Rate Tuning in the Era of Large Language Models. Available online at http://arxiv.org/pdf/2309.08859.

Kalyan, Katikapalli Subramanyam (2024): A survey of GPT-3 family large language models including ChatGPT and GPT-4. In *Natural Language Processing Journal* 6, p. 100048. https://doi.org/10.1016/j.nlp.2023.100048.

Kantrowitz, Alex (2024): ChatGPT's Growth Is Flatlining. In *TheWrap*, 2/16/2024. Available online at https://www.thewrap.com/chatgpt-growth-2024/, checked on 6/23/2024.

Kaufmann, Timo; Weng, Paul; Bengs, Viktor; Hüllermeier, Eyke (2023): A Survey of Reinforcement Learning from Human Feedback. Available online at http://arxiv.org/pdf/2312.14925.

Lehr, R. (1992): Sixteen S-squared over D-squared: a relation for crude sample size estimates. In *Statistics in medicine* 11 (8), pp. 1099–1102. https://doi.org/ 10.1002/sim.4780110811.

Leong, Wei Qi; Ngui, Jian Gang; Susanto, Yosephine; Rengarajan, Hamsawardhini; Sarveswaran, Kengatharaiyer; Tjhi, William Chandra (2023): BHASA: A Holistic Southeast Asian Linguistic and Cultural Evaluation Suite for Large Language Models. Available online at http://arxiv.org/pdf/2309.06085.

Naveed, Humza; Khan, Asad Ullah; Qiu, Shi; Saqib, Muhammad; Anwar, Saeed; Usman, Muhammad et al. (2023): A Comprehensive Overview of Large Language Models. Available online at http://arxiv.org/pdf/2307.06435.

Osváth, Mátyás; Yang, Zijian Győző; Kósa, Karolina (2023): Analyzing Narratives of Patient Experiences: A BERT Topic Modeling Approach. In *ACTA POLYTECH HUNG* 20 (7), pp. 153–171. https://doi.org/10.12700/APH.20.7.2023.7.9.

Otte, Willem M.; Tijdink, Joeri K.; Weerheim, Paul L.; Lamberink, Herm J.; Vinkers, Christiaan H. (2018): Adequate statistical power in clinical trials is associated with the combination of a male first author and a female last author. In *eLife* 7. https://doi.org/10.7554/eLife.34412.

Paaß, Gerhard; Giesselbach, Sven (2023): Foundation Models for Natural Language Processing. Pre-trained Language Models Integrating Media. Cham: Springer Nature (Artificial Intelligence). Available online at https://directory.doabooks.org/handle/20.500.12854/107926.

Pota, Marco; Ventura, Mirko; Fujita, Hamido; Esposito, Massimo (2021): Multilingual evaluation of pre-processing for BERT-based sentiment analysis of tweets. In *Expert Systems with Applications* 181, p. 115119. https://doi.org/10.1016/j.eswa.2021.115119.

Russell, Stuart J.; Norvig, Peter (2022): Artificial intelligence. A modern approach. With assistance of Ming-Wei Chang, Jacob Devlin, Anca Dragan, David Forsyth, Ian Goodfellow, Jitendra Malik et al. Fourth edition, global edition. Boston: Pearson (Always learning).

Available online at https://elibrary.pearson.de/book/99.150005/9781292401171.

Ruxton, Graeme D. (2006): The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. In *Behavioral Ecology* 17 (4), pp. 688–690. https://doi.org/10.1093/beheco/ark016.

Schmidhuber, Jürgen (2015): Deep learning in neural networks: an overview. In *Neural networks : the official journal of the International Neural Network Society* 61, pp. 85–117. https://doi.org/10.1016/j.neunet.2014.09.003.

Smith, Samuel L.; Kindermans, Pieter-Jan; Ying, Chris; Le V, Quoc (2017): Don't Decay the Learning Rate, Increase the Batch Size. Available online at http://arxiv.org/pdf/1711.00489.

Statista (2024): Most used languages online by share of websites 2024 | Statista. Available online at https://www.statista.com/statistics/262946/most-common-languages-on-the-internet/, updated on 12/1/2024, checked on 12/1/2024.

Talaat, Amira Samy (2023): Sentiment analysis classification system using hybrid BERT models. In *J Big Data* 10 (1), pp. 1–18. https://doi.org/10.1186/s40537-023-00781-w.

Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N. et al. (2017): Attention Is All You Need. Available online at http://arxiv.org/pdf/1706.03762.

Web Technology Surveys (2024): Usage Statistics and Market Share of Content Languages for Websites, June 2024. Available online at https://w3techs.com/technologies/overview/content_language, updated on 6/23/2024, checked on 6/23/2024.

Zhang, Jianyu; Bottou, Léon (2024): Fine-tuning with Very Large Dropout. Available online at http://arxiv.org/pdf/2403.00946.

Zhang, Jingzhao; He, Tianxing; Sra, Suvrit; Jadbabaie, Ali (2019): Why gradient clipping accelerates training: A theoretical justification for adaptivity. Available online at http://arxiv.org/pdf/1905.11881.

Zhao, Wayne Xin; Zhou, Kun; Li, Junyi; Tang, Tianyi; Wang, Xiaolei; Hou, Yupeng et al. (2023): A Survey of Large Language Models. Available online at http://arxiv.org/pdf/2303.18223.

Zhao, Yiran; Zhang, Wenxuan; Chen, Guizhen; Kawaguchi, Kenji; Bing, Lidong (2024): How do Large Language Models Handle Multilingualism? Available online at http://arxiv.org/pdf/2402.18815.

Zhu, Q.; Luo, J. (2022): Generative Pre-Trained Transformer for Design Concept Generation: An Exploration. In *Proc. Des. Soc.* 2, pp. 1825–1834. https://doi.org/10.1017/pds.2022.185.

Zhu, Yukun; Kiros, Ryan; Zemel, Richard; Salakhutdinov, Ruslan; Urtasun, Raquel; Torralba, Antonio; Fidler, Sanja (2015): Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. Available online at http://arxiv.org/pdf/1506.06724.