

Does ChatGPT Adapt Itself to the Language Used and the Audience It Implies?

Iglika Nikolova Stoupak
Sens Texte Informatique Histoire
Sorbonne Université, Paris, France
iglika.nikolova-stoupak@
etu.sorbonne-universite.fr

Gaël Lejeune
Sens Texte Informatique Histoire
Sorbonne Université, Paris, France
gael.lejeune@sorbonne-universite.fr

Eva Schaeffer-Lacroix
Sens Texte Informatique Histoire
Sorbonne Université, Paris, France
eva.lacroix@inspe-paris.fr

This paper seeks to quantify and analyse the progress that ChatGPT has made from its GPT-3.5 (2022) to its GPT-4.5 (2025) version when it comes to answering prompts in a selection of differently-resourced languages: English, Bulgarian, Greek, French, Hebrew, Japanese and Russian. Factual correctness, textual quality and an answer's linguistic and cultural independence from an English baseline are evaluated in the process. Each response is marked positively or negatively for each of the three metrics based on a set of defined criteria and careful human-based analysis. In addition, three categories of questions are experimented with: general (e.g. communication assistance or request for jokes), perception-related (e.g. creative writing or explanation of physical processes) and geography-/culture-sensitive (questions in a specific language that address a particular, slightly sensitive topic related to the implied audience e.g. 'Why do French people eat snails?'). As hypothesised, the recent GPT-4.5 version demonstrates significant progress in all evaluated categories, thereby resolving past issues such as decreased textual quality of low-resourced languages and, notably, very limited variety in answers to the same question across languages. The metric 'Independence from the (English) Baseline' receives 80.95% of positive marks in the GPT-4.5 version as opposed to 26.19% for GPT-3.5. Lingering problems include ChatGPT's incomplete ability to generate relevant and culturally-sensitive jokes and poems.

Keywords: ChatGPT, GPT-3.5, GPT-4.5, multilingualism, bias

1. Introduction

OpenAI's ChatGPT barely needs introduction as of today. Appreciated by AI specialists and the general public alike, the chatbot assists internet users from all around the world in a number of tasks from social advice to academic writing and coding. It is based on the GPT (Generative Pre-trained Transformer) model, which achieves state-of-the-art performance

<https://doi.org/10.47810/JCLIB.1.2025.02>

© 2025 Institute for Bulgarian Language, Bulgarian Academy of Sciences, Department of Computational Linguistics
Published under a Creative Commons Attribution 4.0 International Licence (CC BY 4.0).

in a number of natural language processing (NLP) tasks. ChatGPT is user-friendly with its strong zero-shot capabilities and, due to the incorporation of Reinforcement learning from Human Feedback in its training stages, its output highly resembles human language. On the other hand, ChatGPT’s limitations as expressed in research since its release in 2022 include lower functionality in low-resourced languages as well as the presence of sometimes concerning biases in output. Since the system’s conception, serious steps have been taken to eliminate these problems as well as to increase ChatGPT’s overall quality.

The GPT-4 family, trained on a larger dataset than the initial GPT-3.5, comes with an increased general ability to understand and generate text (Caramancion 2023). It has been noted to provide improved answers to math- and logic-related questions (Plevris, Papazafeiropoulos, and Rios 2023). Whilst the GPT-3 family is, in general, competitive with state-of-the-art language models fine-tuned for specific tasks, its GPT-4 descendant outperforms most state-of-the-art systems (Brown et al. 2020; OpenAI et al. 2024a). Similarly, the latter’s scores on a simulated bar exam fall within the top 10% for human test takers, compared to the bottom 10% for GPT-3.5 (OpenAI et al. 2024a).

In light of this raising of the bar, the time might be right to move beyond the initial concerns of textual errors in low-resourced languages and outright harmful content and extend the focus onto the system’s more subtle malleability. Language implies underlying culture, geography and history and, ideally, ChatGPT should be expected to respond in accordance with this user background. For instance, lengthy and English-based formulas of politeness might not sound natural to a Japanese speaker, passive constructions may impede reader comprehension in Slavic languages, and a translation or explanation of an already French-based culinary term may be rather unnecessary for a French audience.

In an attempt to test ChatGPT’s success and progress in adapting to its multilingual audience, a selection of languages with significantly different resource availability was opted for. See Table 1 for a rough overview of each selected language’s online prevalence as based on the existent number of Wikipedia pages written in it¹. A variety of geographical and cultural characteristics related to the languages’ native speakers was also sought. Two Slavic languages were deliberately opted for: Russian and Bulgarian. The fact that the former is significantly more highly-resourced than the latter allows for the detection of possible interference as well as for conclusions to be drawn about the relative importance of language resourcedness versus language similarity within the proposed experiments.

In order to evaluate ChatGPT’s full up-to-date progress, we will compare the performance of its first and most recent versions; respectively, GPT-3.5 and GPT-4.5. As per its system card, GPT-4.5 builds upon the previous model, GPT-4o, while being more general-purpose in nature. Its underlined strenghts include ‘alignment with user intent’ and ‘improved emotional intelligence’ (OpenAI 2025). The model is currently limited to Plus and Pro subscription plans and a quota of around 50 messages per week is attributed in the former.

Whilst the focus of the current study is specifically ChatGPT’s ability to adapt to its user as implied by the prompt’s language, we are using the opportunity to coincidentally

1 as per https://meta.wikimedia.org/wiki/List_of_Wikipedias

Language Code	Number of Pages
EN	62,907,668
FR	13,440,180
RU	8,246,967
JA	4,279,143
HE	1,579,325
EL	716,967
BG	684,829

Table 1
Number of Wikipedia pages per utilised language

evaluate the model’s performance and progress in more traditional aspects, such as factual correctness and textual quality, especially given evidence that low-resourced languages have been associated with output of lower quality due to an involved process of translation (Zhang et al. 2023).

Based on ChatGPT’s specifications as per OpenAI and relevant academic research, it is hypothesised that ChatGPT’s recent version will perform significantly better across languages and across the different question categories. The following specific research questions are brought forward in relation to the experiments’ results:

1. Is subordinate multilingualism (characterised with prior translation) less prominent in the case of GPT-4.5?
2. Does ChatGPT’s GPT-4.5 version provide a higher proportion of language-/culture-sensitive answers? If yes, to what extent?
3. Does ChatGPT provide a different degree of malleability with regard to the different defined question categories?
4. Are there discernible tendencies in relation to specific languages (e.g. as characterised by language family or script)? In particular, is there interference from the higher- to the lower-resourced Slavic language involved?

2. Background

Although to our knowledge, no extensive research has addressed the specific ability of ChatGPT to adapt itself to its audience as demonstrated by the prompt language, it is worth examining the chatbot’s performance in two directly related aspects: biases and multilingual output.

2.1 ChatGPT and Biases

Since the launch of ChatGPT, ethical issues have been raised in relation to detected biases based on gender, race, religion, occupation, etc., likely caused by existing biases in the

utilised training data. For instance, Plevris et al (2023) discover that the word ‘Black’ is associated with a consistently low sentiment in the model’s output. Deshpande et al (2023) note the model’s increased tendency to incorporate stereotypes when prompted to act as a specific persona. In their extensive study on ChatGPT’s robustness and ethics, Vidhya et al (2023) examine the phenomenon of ‘jailbreak’ i.e. the model’s tendency to bypass ethical norms as a result of specific prompt engineering techniques. Possibly controversial output is detected in several languages, for instance in relation to contested territory (Vidhya et al. 2023). Rozado discusses political and demographic biases as displayed by ChatGPT and other large language models (LLMs), such as their failure to always flag negative content accordingly (Rozado 2023).

The OpenAI team set out to work towards removing the existing biases ‘in a holistic manner’ (Brown et al. 2020). Their measures, manifested in the GPT-4 model family, include alignment to human preferences during post-training (OpenAI et al. 2024b), training the model for refusals, and the assembly of a red team of experts that monitors the model’s activity (OpenAI et al. 2024a). However, biases and the bypassing of ethical norms continue to represent a problem. For example, in an instance of jailbreak, ChatGPT is led to propose an antisemitic comment that would not trigger flagging on Twitter (OpenAI et al. 2024a). To go further, the activity of the red team itself, which is mostly made up of English speakers from Western countries, may lead to the favouring of specific opinions and worldviews (OpenAI et al. 2024a).

In their detailed study, Puttaparthi et al (2023) research the effect of multilingual wrapping of prompts on the probability of ChatGPT jailbreak. 5.07% of questions asked in a single language lead to jailbreak (none of them being in English). Out of the remaining questions, a further 3.21% cause the model to fall into the trap when reformulated in a mixture of languages. In a final experiment, the yet remaining questions are repeated, instructing ChatGPT to provide output in a language not present in the prompt, leading to 1.61% of jailbreak. Furthermore, the phenomenon’s probability increases significantly when prompt injections² are used. Puttaparthi et al also note that the process of multilingual wrapping tends to impede the model’s comprehension and leads to output of limited quality.

2.2 ChatGPT and Multilingualism

ChatGPT’s multilingual abilities and the limitations therein also come as a frequent subject of discussion. The GPT-3.5 model is trained almost exclusively on high-quality English language data from the Common Crawl, text in other languages accounting for only 7% of the data (Brown et al. 2020). Still, ChatGPT has been noted to outperform earlier LLMs such as T5 and BERT across a variety of NLP tasks when diverse languages, including low-resourced ones, are concerned (Lai et al. 2023). However, the model’s performance sharply deteriorates in the presence of extremely low-resourced languages, such as Buginese. It has also been noted that while ChatGPT exhibits comprehension in some rare languages, it struggles to

² a cybersecurity exploit that seeks to confuse a model by providing it with both a legitimate prompt and a request to ignore it and offer specific, different output

identify the language itself (Bang et al. 2023). Peng et al (2023) discover a higher number of hallucinations in machine translation (MT) output when low-resourced languages as well as distant language pairs are involved. The model’s limitations in multilingual performance also impact tasks that are tightly associated with ethics. For instance, Das et al (2023) evaluate ChatGPT’s ability to detect hate speech within input in 11 discrete languages and detect weaknesses in the model’s identification of irony as well as the differentiation of protected target groups within non-English languages.

As is the case with the majority of the model’s functions, ChatGPT demonstrates significantly increased multilingual performance in its GPT-4 version. Manakhimova et al (2023) test the model’s performance in MT in the following language pairs: German-English, English-German and English-Russian. Whilst the pair that involves Russian (a language linguistically remote from the other two) performs worst and issues such as the interpretation of idioms are still detected, the results are comparable to those of state-of-the-art systems. Jiao et al (2023) research ChatGPT’s translation ability in 101 languages based on both its GPT-3.5 and GPT-4 versions. When it comes to the former, the gap between high- and low-resourced languages is claimed to be drastic, although it can be partially mitigated through techniques such as pivot-prompting (i.e. explicitly asking the model to translate into a higher-resourced language as an additional step). In contrast, the more recent GPT version is associated with strikingly higher BLUE scores, and its performance is deemed to be of sufficiently high quality even when low-resourced languages are concerned. With GPT-4, languages such as Latvian, Welsh and Swahili show strong performance on the MMLU benchmark of multiple-choice questions on 57 subjects (OpenAI et al. 2024a). In turn, GPT-4o demonstrates improved reading comprehension and reasoning abilities in historically underrepresented languages, significantly narrowing their gap with English (OpenAI et al. 2024b).

A specific study that is worth mentioning in view of the current research is Zhang et al’s (2023) comprehensive investigation of the type of bilingualism exhibited by ChatGPT when it provides non-English output (assuming that English is the model’s ‘native language’). They bring forward the following terms as drafted by Marcos (1976): coordinate bilingualism (wherein one’s lexicons for each language are associated with discrete mental images) versus subordinate bilingualism (wherein translation occurs into one’s main language prior to textual production). They experiment with three prompt categories which imply different degrees of impact of the language involved: Reasoning, Knowledge Access, and Articulation. The last category is defined as ‘translation variant’; that is to say, different output in terms of content is to be expected in different languages³. Produced answers in non-English languages in this category are discovered to be very similar to their English counterparts, as per the cosine distance of their BERT embeddings (following initial automatic translation of the non-English text). Overall, ChatGPT is concluded to exhibit a mixture of coordinate and subordinate bilingualism. The authors also note that as a result of underlying translation, ‘an errorprone process’, the accuracy of output achieved through subordinate bilingualism suffers deterioration (Zhang et al. 2023). The question remains of whether later versions

3 An example is the composition of a cover letter, which is affected by linguistic and societal norms.

of ChatGPT have led to a shift or even a qualitative change in the described multilingual abilities.

3. Methods

ChatGPT as per its GPT-3.5 and GPT-4.5 versions was asked to provide answers to a series of prompts in a zero-shot setting. A new session was started following each output in order to avoid the system gaining knowledge of the fact that an academic experiment was being performed and/or that its interlocutor was using multiple languages. ‘Incognito’ mode was not made use of for the following reason: experiments for the current research commenced in 2023, when ChatGPT did not typically demonstrate any knowledge of the user’s background or prior conversations once a new conversation was started. The system was accessed via its web-based direct chat interface, through a ‘Plus’ subscription plan.

3.1 Question Types

Different types of questions were employed (see Table 3.1), calling for different degrees of sensitivity to the language and culture at hand within the output. Coincidentally, different tasks and language registers were also implied.

Question type 1 involves general, frequently encountered questions, linked to help with writing and planning, social interaction and entertainment. Social norms and accepted levels of politeness are dependent upon historical phenomena, such as a country’s relationships with industrialisation and occupational self-direction (Schooler 1996). Humour in turn has been proven to be both universal and culture-specific. For instance, Jiang et al (2019) note that Eastern cultures are less receptive to aggressive humour compared to Western ones and that, in particular, humour is often used as a coping social device in Japan.

Type 2, broadly named ‘perception-related’, touches on aspects of language that are directly related to one’s perception of the world, including spacial orientation, climatic specificity, agency, colour perception and gender representation. Various theories exist on the question of whether or to what extent language determines thought. As per the Sapir-Whorf hypothesis, language at the very least mediates or influences perception (Fulga 2012). For instance, colours interrelate with metaphorical extensions (van Leeuwen 2010), and the direction of writing in one’s native language affects the response to and production of movement (Boroditsky 2011). Furthermore, the linguistic representation of time has been claimed to strongly affect one’s thought process (Fulga 2012).

Finally, question type 3 includes topics and concerns that are specifically relevant to an audience from a particular country or countries. The topics were selected to be emotionally charged to the point of being slightly controversial whilst not reaching the severity point of testing the system for harmful biases. For example, a common misconception exists that the weather in Russia is very cold, whilst this is only true for a limited area within the country.

For question types 1 and 2, all prompts were provided in each of the seven investigated languages. In contrast, each type 3 question was presented solely in English and the particular language it concerns. English, being the main language of ChatGPT’s training and proven

Question	Expectations
1. General 1.1. Please write an email telling my dentist that I cannot make it to my appointment and I will have to reschedule it. 1.2. I am very shy. Can you give me advice about how I can meet new people? 1.3. Tell me a joke.	consideration of culture-specific etiquette, societal norms, degrees of sharing and humour
2. Perception-related 2.1. Write a 4-line poem about autumn using names of colours in it. 2.2. Write a 2-sentence piece of creative writing. In sentence 1, say how quickly the summer passed. In sentence 2, describe the autumn using a metaphor or simile. 2.3. Describe how a mechanical clock works (up to 200 words).	consideration of culture- and language-specific artistic conventions, climate specificity, colour perception, gender-related imagery, rhyme, suitable verbs and verb forms (e.g. passive vs active; transitive vs intransitive)
3. Geography-/Culture-sensitive 3.1. Is Bulgarian the oldest Slavic language? 3.2. Did the ancient Greeks speak Greek? 3.3. Why do French people eat snails? 3.4. Why does Israel participate in Eurovision? 3.5. Is it hard to work for a Japanese company? 3.6. What is the weather like in Russia?	assumption of a topic’s sensitivity; different degrees of expected knowledge; different levels of detail

Table 3.1
Questions and expected model behaviour

to evoke responses of highest quality, was taken as the baseline language. Three answers per language were recorded for question 1.3., which solicited jokes from the chatbot, as the output was of very short length (thus affecting statistical analysis) and tended to differ significantly in discrete sessions.

3.2 Evaluation Criteria

ChatGPT’s output was evaluated in view of three criteria: ‘Factual Correctness’, ‘Textual Quality’ and ‘Independence from the English Baseline’; the third one being the focus of the current study. For a detailed explanation of the criteria, please refer to Table 3.2.

Evaluation was based on human analysis carried out by the authors (who are professionals in linguistics). Importantly, in the events of doubt as to a given answer’s quality and naturalness, native speakers, who are also language teachers of the language in question, were consulted. Each response was marked positively or negatively (‘Yes’ versus ‘No’) for the three evaluation criteria. This binary approach was judged to be preferable to one involving a larger scale given the focus on human-based and partly subjective evaluation (in contrast and in complement to the one offered by Zhang et al. (2023)). In the case of question 1.3., the mode score for the three output jokes was considered.

Category	Criteria
Factual correctness	the answer corresponds to the prompt ^a ; the language of the answer corresponds to the language of the prompt; if relevant, the presented objective information is correct; a single failure in the fulfillment of these criteria leads to a negative mark
Textual quality	the output text contains fewer than 0.89 mistakes per 100 words; mistakes include misspellings, wrong grammar, wrong word/collocation choice, unnatural calque, and words in a foreign language
Independence from baseline (non-English text only)	the text is significantly different from the English-language version and the differences make it more appropriate for the audience implied by the prompt’s language; isolated borrowings/calques do not lead to a negative mark when the majority of the text is different from the English baseline

Table 3.2

Evaluation categories and criteria

^a In the case of jokes, output that cannot easily be viewed as an attempt at humour receives a negative mark.

‘Textual Quality’ was calculated in the following manner:

1. All mistakes in the model’s output were identified.
2. The number of words in each answer was calculated (for Japanese text, which contains no spaces, the python library *tinysegmenter*⁴ was deployed).
3. For each answer, the number of mistakes per 100 words was calculated.
4. A distribution-based threshold of 0.89 was determined as the value beyond which answers received a negative mark for ‘Textual Quality’⁵.

⁴ <https://pypi.org/project/tinysegmenter/>

⁵ This value was selected as it leads to 75% of the answers being marked positively.

English-language output was also evaluated for ‘Factual Correctness’ and ‘Textual Quality’, but the results did not take part in the majority of subsequent statistical analyses⁶ for the following reasons: English has the role of baseline and the study’s key focus is on output’s independence from the baseline; in addition, no concerns were detected in the language concerning the two applicable criteria.

Additional qualitative observations, such as the exact nature of each output’s strengths and weaknesses, were also taken note of in the evaluation process.

For the evaluation results as well as ChatGPT’s full output for the language-/culture-sensitive questions, please refer to Appendix A (version GPT-3.5) and Appendix B (version GPT-4.5). For the evaluation and output of all sets of questions, please access the following GitHub repository: https://github.com/iglika88/ChatGPT_language_audience_adaptation.

4. Results

Figure 1 shows the percentage of ‘Yes’ scores per evaluation measure for the two investigated versions of ChatGPT. Whilst both ‘Factual Correctness’ and ‘Textual Quality’ reach 100% for GPT-4.5, the category ‘Independence from the Baseline’ is associated with the largest leap in the model’s newer version: from 26.19% to 80.95%.

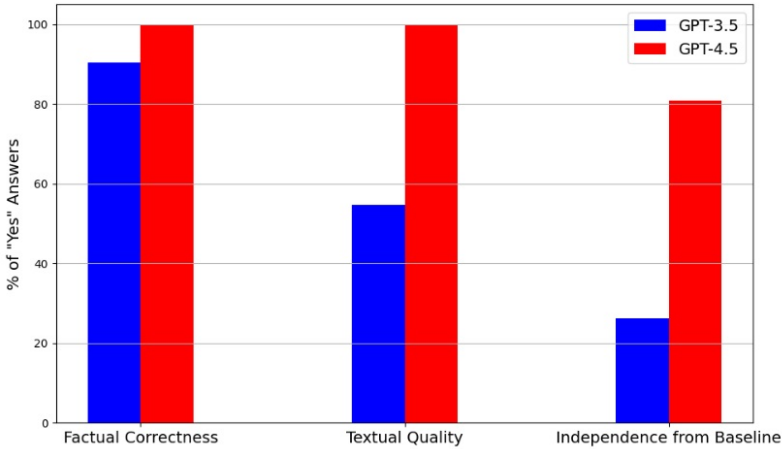


Figure 1
Percentage of ‘Yes’ scores per evaluation measure

Figures 2, 3 and 4 present the same information by language. ‘Factual Correctness’ already reaches maximum value in GPT’s 3.5 version in relation to Bulgarian and French. In contrast, ‘Textual Quality’ is lowest for Bulgarian (i.e. the lowest-resourced language). Bulgarian, Greek, French and Russian are the languages with weakest scores for ‘Independence

⁶ An exception is the calculation of a correlation between textual quality and language resourcedness.

from the Baseline’ for the GPT-3.5 version (14.29%) and Russian for the GPT-4.5 version (57.14%). The Hebrew language shows the smallest improvement in the category (57.14% to 71.43%), passing from the strongest to the second weakest position. Such shifts in rank may speak of a qualitative difference in the system’s performance between the two investigated versions. This hypothesis is also supported by the fact that, interestingly, it is Bulgarian and Japanese that score highest for the category in the GPT-4.5 version while being, respectively, the least-resourced and most different from English languages, characteristics that have been associated with reduced performance in ChatGPT.

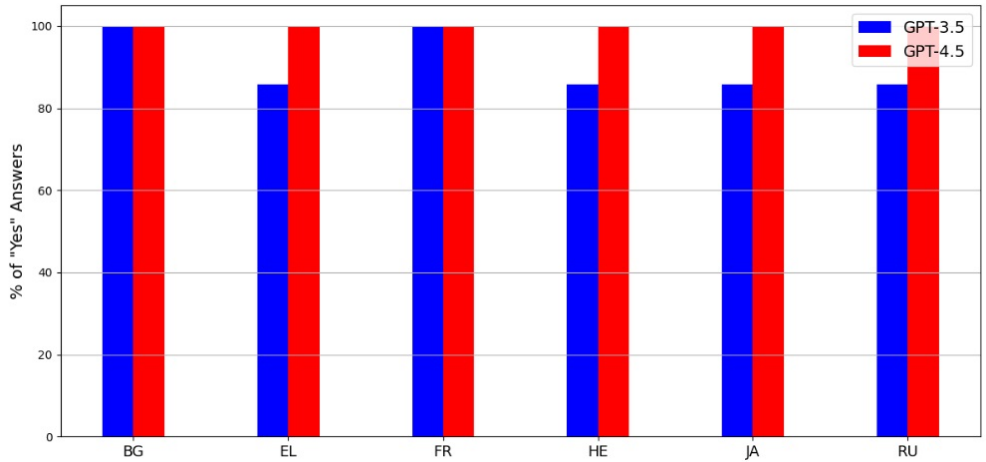


Figure 2
Percentage of ‘Yes’ scores for ‘Factual Correctness’ per language

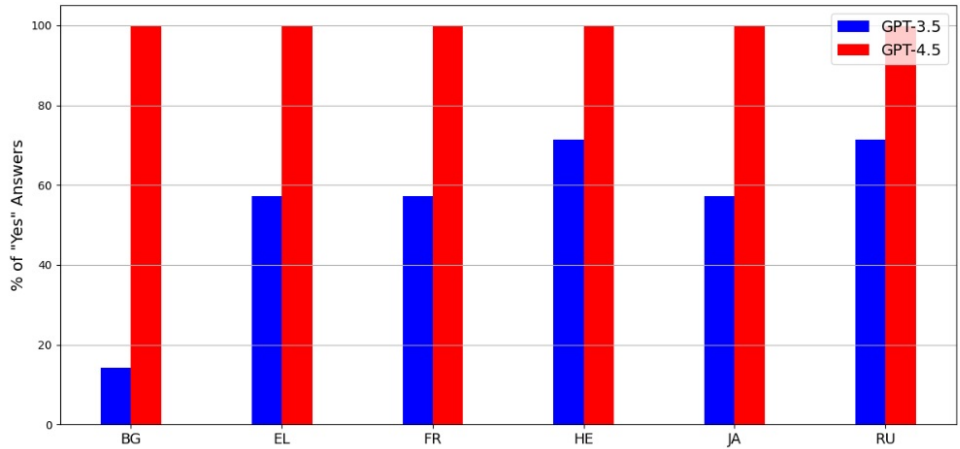


Figure 3
Percentage of ‘Yes’ scores for ‘Textual Quality’ per language

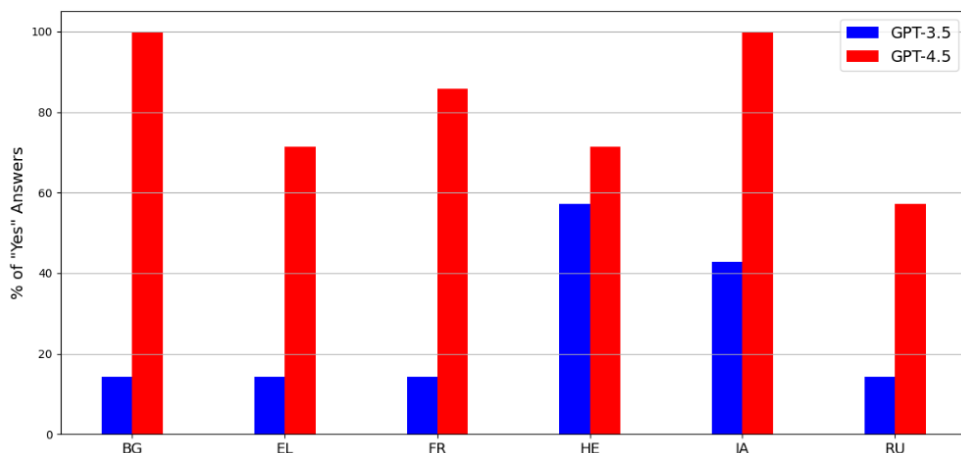


Figure 4

Percentage of 'Yes' scores for 'Independence from Baseline' per language

A further focus on each of the three different question types as defined in Section 3.1 reveals that perception-related questions are most challenging to GPT-4.5 (Figure 5). It is interestingly namely this question type that is marked with the least progress from GPT-3.5. Please refer to Appendix C for a further breakdown of the models' scores by question type and by language. ChatGPT's current version performs weakest in Greek, French and Hebrew for 'general' questions and in Greek, Hebrew and Russian for 'perception-based' ones.

The potential correlation between language resourcedness (as per the proxy of number of Wikipedia pages) and 'Textual Quality' as well as 'Independence from the Baseline' was subsequently examined. A Pearson coefficient of 0.70 was revealed in relation to the first measure in GPT's 3.5 version, denoting a significant positive relationship. In contrast, no correlation was detected with independence from the baseline. For the associated graphs, please refer to Appendix C.

5. Discussion

Figure 6 provides a specific example of juxtaposed output for the same question (2.2) in the English versus Bulgarian languages per the two examined versions of ChatGPT. The quality of GPT-3.5-generated text is perfect in English, whilst the Bulgarian output contains examples of wrong word choice (*отсвирвания*, a noun from an informal verb meaning 'to stop an activity' or 'to chase away someone') and unnatural collocation (*слънчеви забавления* 'sunny entertainment'). The featured figurative language and imagery are very similar in the two languages (summer slipping through one's fingers; autumn compared to a painter's canvas/brushstrokes). In contrast, the Bulgarian text provided by version GPT-4.5 is error-free. Also, although some common elements are detectable in the two languages' content (the 'painting' linguistic field, the colour 'gold'), a number of differences can be

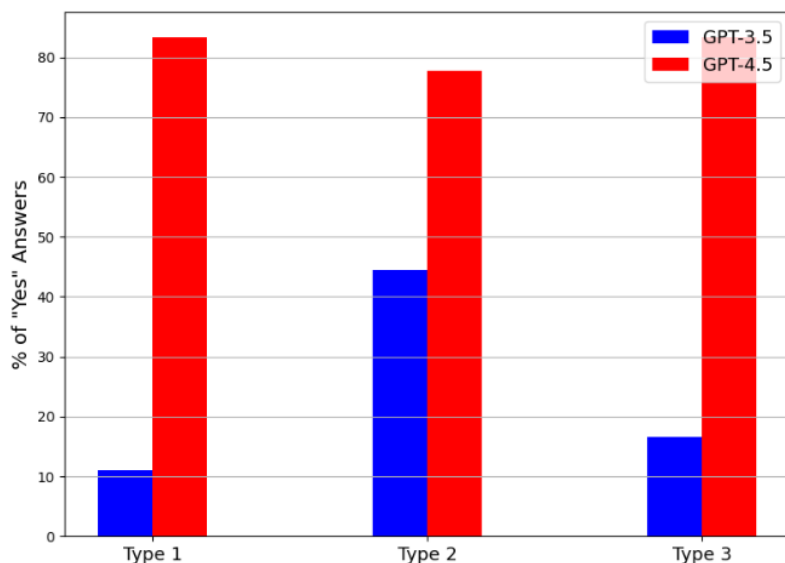


Figure 5

Percentage of 'Yes' scores for 'Independence from Baseline' per question type

	GPT-3.5	GPT-4.5
EN	Summer slipped away like sand through eager fingers, leaving only fading echoes of laughter and warmth. Autumn arrived, draped in a tapestry of fiery leaves, like a painter's brushstrokes on the canvas of the world.	Summer slipped away like a secret whispered too softly to catch. Autumn arrived, painting the world in shades of amber and gold, like a poet quietly setting words ablaze.
BG	Лятото измина като миг между пръстите, оставяйки зад себе си само приглушени отсвирвания на слънчеви забавления. Есента дойде като художествено платно, покрито с листа, сякаш четка на художник ги рони, създавайки картина от топли оттенъци.	Лятото отлетя като хвърчило, което вятърът отнесе отвъд хоризонта, преди още да съм се научил да го задържам. Есента пристъпи тихо, като художник с шепя злато и мед, готов да преобрази света с последните си щрихи.

Figure 6

Juxtaposition of the English versus Bulgarian output for question 2.2 as per the two examined versions of ChatGPT

pointed out. Summer is compared to 'a secret' in English and to 'a kite' in Bulgarian, the latter evoking childhood imagery that is common in Bulgarian songs and poetry. There is an instance of the first person in Bulgarian (*преди още да съм се научил да го задържам*

‘before I have even learnt to use it’). Finally, the Bulgarian text is significantly longer than its English counterpart, which occurs when the two languages are naturally used.

Let us address the research questions as stated in Section 1.

1. GPT-4.5’s output scores perfectly for ‘Textual quality’, thereby implying that there has been no loss in quality as a result of translation. Other reasons to deduce the absence of a translation process in GPT-4.5 include: the use of natural formulas of politeness for all languages in question 1.1 (e.g. (BG) *с уважение* ‘with respect’; (FR) *je vous remercie par avance de votre compréhension* ‘thank you in advance for your understanding’); the general absence of calques, false friends and foreign words, which were often encountered in the GPT-3.5 version (e.g. (EL) *ξεκινήστε μικρά*, calque from ‘to start small’); and the absence of unnatural instances of the passive voice (e.g. (BG) *когато пружината се навива* lit. ‘when the spring winds itself’ is used in question 2.3. in place of the non-native-sounding *след като бъде намотана* ‘after it is wound’ found in GPT-3.5).

The non-English jokes presented by GPT-4.5 are of a particularly higher quality compared to their GPT-3.5 counterparts, which are saturated with incomprehensible ‘puns’, likely resulting from literal translation from English (e.g. a Japanese-language joke says that a tomato won a race *ソースだったから!* ‘because it was a sauce!’; revealing a possible underlying ‘catch-up/ketchup’ pun). Moreover, some poems in GPT’s 3.5 version contain rhymes traceable back to an intermediary English text; for instance, (RU) *дня - игру* ‘day - play’.

2. While the GPT-4.5 model still contains negative scores for the category ‘Independence from the baseline’ (in particular, when it comes to the ‘perception-based’ question type), it is associated with a remarkable increase of 54.76% from GPT-3.5.

Several promising instances of sensitivity to the user as per their language are already noted in relation to the GPT-3.5 version. The colours mentioned in the Japanese answer to question 2.1 are very specific and contain subtle differences, possibly reflecting the rich haiku culture: 褐色, ‘brown’, 深紅 ‘deep red’ and 紅 ‘red’. The Japanese passive voice, which so much as caused linguists to reformulate the universal characteristics of grammar (Ishizuka 2012), is used effectively in the answer to question 2.3: 構成されています ‘has been constructed’ replaces the simple ‘functions’, encountered in the English baseline. Finally, the Hebrew answer to the same question is the only one that does not include the phrase ‘hours, minutes, and seconds’, possibly reflecting on the fact that these divisions of time were not found in the language until late in its development (Kogan et al. 2007).

Within GPT-4.5, answers to the same questions generally differ significantly between languages, the tendency being most clear within the third question type. For instance, the Hebrew-centred question includes information in Hebrew that is absent in English, such as the years when Israel won the Eurovision contest. Similarly, in Japanese, unlike the English baseline, the question about the challenging aspects of working for a Japanese company is not met with reference to a ‘language barrier’, implying an assumed fluent interlocutor. Question 2.1., which requests a short poem that includes the names of colours, receives very specific output in Russian due to the fact that the word for ‘colours’ and ‘flowers’ is the same in the genitive case (*цветов*): inventively, names of both colours and flowers are included (*золотые* ‘golden’, *седой* ‘grey’; *хризантемы* ‘chrysanthemums’, *розы* ‘roses’, *астры* ‘asters’).

3. ‘General’ questions score worst for ‘Independence from the baseline’ in GPT-3.5 (in particular, the composition of jokes and social advice). In contrast, ‘perception-based’ questions are overall most challenging for GPT-4.5. Question 2.2, whose output is a short piece of creative writing, and question 1.3, whose output are jokes, receive the lowest scores (50%).

4. In the GPT-3.5 version, Bulgarian, which is the lowest-resourced of the examined languages, and Japanese, which is the most remote one from English in terms of language family and the nature of the alphabet, have consistently low scores across different questions for both ‘Textual quality’ and ‘Independence from the baseline’ (see Figures 2 and 3). In contrast, these languages score highest in the GPT-4.5 version (see Figure 4).

There are reasons to believe that in GPT-3.5 Bulgarian, a Slavic language, is influenced by Russian, another Slavic language that is significantly higher-resourced. Russian words that don’t exist in Bulgarian are included in Bulgarian text (e.g. *начнете*, the imperative form of ‘to begin’, in question 1.2.). The importance of language resourcedness for GPT-3.5’s output quality is demonstrated by the fact that the passive voice, equally unnatural in both Slavic languages, is present in the Bulgarian answer to question 2.3. but absent from its Russian counterpart. Similarly, in question 2.2. autumn is described as ‘blossoming’ (*расцвела*) in Russian, imagery that can be seen as fitting the noun’s feminine gender in the language. In contrast, feminine imagery is not found within the Bulgarian output (even though the noun is also feminine). ChatGPT may be going further and assuming cultural similarity based on language family. One of the Russian-language jokes approaches the topic of space travel, which can be considered as important for an implied Russian-speaking audience; interestingly, one of the Bulgarian-language jokes (and no jokes in other languages) also features this topic despite its lack of comparable relevance for the implied audience.

When it comes to the GPT-4.5 version, a possible instance of interference based on the proximity of the two Slavic languages is detected in question 3.6, whose topic is linked to Russia. There, the English prompt leads to output that is in Bulgarian rather than English. This could easily be explained by the fact that the researcher’s operating system was set in Bulgarian and that the output relied on real-time information (the current weather in Russia). However, the phenomenon did not recur during an additional experiment, where the weather in a variety of other geographical locations was requested. Therefore, the system might assume an especially direct link between Bulgarian and Russian audiences, even when non-linguistic aspects of the output (such as implied geography) are concerned.

6. Conclusion and Future Work

As has been noted by previous research, there are reasons to believe that ChatGPT as per its GPT-3.5 version applied translation when composing non-English text. Associated problems included: a high gap in performance in relation to low-resourced languages and languages that are very dissimilar to English, interference of related higher-resourced languages, use of unnatural calques, and uniformity of answers to the same question when provided in different languages. These problems largely disappear with GPT-4.5. This significant progress speaks of as much as a qualitative change that one may describe as a move from subordinate

to coordinate multilingualism, whether it came about as a result of additional training data or deliberate modification in the model’s architecture or training. This being said, ChatGPT still has room for progress: the time is yet to come when it will offer exquisite haikus or side-splitting jokes.

The present study may serve as a benchmark for further evaluation of the progress of GPT models. The involvement of additional languages would clearly help to confirm the validity of the conclusions reached. Other potential avenues for extension include experiments with one-shot and few-shot settings as well as with distinct LLMs.

Limitations

Firstly, it is worth noting that within experiments, a general assumption has been made that ChatGPT’s users compose prompts in their (closely) native language, which is not always the case. Also, while clear definitions of the evaluation categories are sought, a significant degree of subjectivity remains at the attribution of marks, in particular in relation to the category ‘Independence from the baseline’, as it involves judgment of any differences’ significance when one has in mind the culture and language at hand. The attempt to provide uniform measures across highly different languages also comes with limitations; for instance, there are alternative methods for counting the number of words in Japanese text. Finally, we will mention that this is a temporal snapshot of ChatGPT’s progress that may not be indicative of the rate and nature of the model’s future development.

Acknowledgments

We would like to thank language teachers Katerina Rizopoulou (Greek), Ori Ziring (Hebrew), and Hitoshi Tsunekawa (Japanese) for their enthusiasm and help at evaluating the quality and naturalness of ChatGPT’s responses in the respective languages.

References

- Bang, Yejin, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.
- Boroditsky, Lera. 2011. How language shapes thought: the languages we speak affect our perceptions of the world. <https://www.scientificamerican.com/article/how-language-shapes-thought/>. Accessed: 27/11/2023.
- Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Caramancion, Kevin Matthe. 2023. News Verifiers Showdown: A Comparative Performance Evaluation of ChatGPT 3.5, ChatGPT 4.0, Bing AI, and Bard in News Fact-Checking. *arXiv preprint arXiv:2306.17176*.
- Das, Mithun, Saurabh Kumar Pandey, and Animesh Mukherjee. 2023. Evaluating ChatGPT’s Performance for Multilingual and Emoji-Based Hate Speech Detection.

- Deshpande, Ameet, Vishvak Murahari, Rajpurohit Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in ChatGPT: Analyzing Persona-Assigned Language Models. *ArXiv*, abs/2304.05335.
- Fulga, Angelica. 2012. Language and the perception of space, motion and time.
- Ishizuka, Tomoko. 2012. *The Passive in Japanese: A Cartographic Minimalist Approach*. John Benjamins Publishing Company.
- Jiang, Tonglin, Hao Li, and Yubo Hou. 2019. Cultural differences in humor perception, usage, and implications. *Frontiers in Psychology*, 10.
- Jiao, Wenxiang, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is ChatGPT a Good Translator? Yes With GPT-4 As The Engine.
- Kogan, Leonid E., Natalia Koslova, Sergey Loesov, and Serguei Tishchenko. 2007. *Babel und Bibel 3: Annual of Ancient Near Eastern, Old Testament and Semitic Studies*. Penn State University Press.
- Lai, Viet Dac, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. *arXiv preprint arXiv:2304.05613*.
- van Leeuwen, Theo. 2010. *The Language of Colour: An Introduction*, 1 edition. Routledge.
- Manakhimova, Shushen, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian M"oller. 2023. Linguistically Motivated Evaluation of the 2023 State-of-the-Art Machine Translation: Can ChatGPT Outperform NMT? In *Proceedings of the Eighth Conference on Machine Translation*, pages 224–245, Association for Computational Linguistics, Singapore.
- Marcos, Luis R. 1976. Linguistic dimensions in the bilingual patient. *American Journal of Psychoanalysis*, 4(36):347–354.
- OpenAI. 2025. OpenAI GPT-4.5 System Card. *OpenAI*.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, and ... 2024a. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.
- OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, and ... 2024b. GPT-4o System Card. *arXiv preprint arXiv:2410.21276*.
- Peng, Keqin, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards Making the Most of ChatGPT for Machine Translation.
- Plevris, Vagelis, George Papazafeiropoulos, and Alejandro Jiménez Rios. 2023. Chatbots Put to the Test in Math and Logic Problems: A Comparison and Assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard. *AI*, 4(4):949–969.
- Puttaparthi, Poorna Chander Reddy, Soham Sanjay Deo, Hakan Gul, Yiming Tang, Weiye Shang, and Zhe Yu. 2023. Comprehensive Evaluation of ChatGPT Reliability Through Multilingual Inquiries. *arXiv preprint arXiv:2312.10524*.
- Rozado, David. 2023. Danger in the machine: The perils of political and demographic biases embedded in ai systems. *New Zealand Institute of Skills and Technology*.
- Schooler, Carmi. 1996. Cultural and social-structural explanations of cross-national psychological differences. *Annual Review of Sociology*, 22:323–349.
- Vidhya, N. G., D. Devi, A. Nithya, and T. Manju. 2023. Prognosis of exploration on Chat GPT with artificial intelligence ethics. *Brazilian Journal of Science*, 2(9):60–69.
- Zhang, Xiang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't Trust ChatGPT When Your Question Is Not in English: A Study of Multilingual Abilities and Types of LLMs. *arXiv preprint arXiv:2305.16339*.

Appendices

A. Full Answers and Scores (Question 3): GPT-3.5

Language mistakes are denoted between asterisk (*) symbols.

3. Geography-/Culture-Sensitive Questions		
EN		
3.1. Is Bulgarian the oldest Slavic language?		
<p>Bulgarian is one of the oldest Slavic languages with a rich history and significant antiquity within the Slavic language group. However, determining the absolute oldest Slavic language is a complex task.</p> <p>The Slavic languages are part of the larger Indo-European language family. Proto-Slavic, the hypothetical common ancestor of all Slavic languages, is believed to have existed around 1,500 to 2,000 years ago, but it wasn't a written language. Over time, it evolved into various Slavic languages, including Bulgarian.</p> <p>Bulgarian, like other Slavic languages, has undergone its own unique evolution, influenced by historical, cultural, and geographical factors. The development of the Bulgarian language can be traced back to Old Bulgarian (Old Church Slavonic) and the Glagolitic and Cyrillic scripts.</p> <p>It's essential to acknowledge that all Slavic languages have ancient roots and historical significance, and pinpointing a single "oldest" Slavic language is a challenging task due to the gradual and continuous evolution of languages over time. Additionally, languages are constantly changing, adapting, and diversifying, making it difficult to pinpoint an exact point of origin.</p>		
Factual correctness	Textual quality	Independence from baseline
Y	Y	N/A
BG		
3.1. Българският най-старият славянски език ли е?		
<p>Българският е един от най-старите славянски езици с богата история и *значителна древност* в рамките на славянската езикова група. Въпреки това е трудно със сигурност да се каже кой славянски език е най-стар.</p> <p>Славянските езици са част от по-голямото семейство на индоевропейските езици.</p> <p>Протославянският, хипотетичният общ предшественик на всички славянски езици, се предполага, че е съществувал преди около 1 500 до 2 000 години, въпреки че не е бил писмен език. С течение на времето той еволюира в различни славянски езици, включително и българския.</p> <p>Българският, като и другите славянски езици, претърпява своя собствена *уникална еволюция*, влиянието на исторически, културни и географски фактори. Развитието на българския език може да се проследи до Старобългарския (Старославянски) език и до глаголицата и кирилицата.</p>		

Важно е да се отбележи, че всички славянски езици имат древни корени и историческо значение, и е трудно да се определи един "най-стар" славянски език поради постепенната и продължителна еволюция на езиците с времето. Освен това, езиците постоянно се променят, адаптират и разнообразяват, което затруднява определянето на точна точка на произход.

Factual correctness	Textual quality	Independence from baseline
Y	N	N

EN

3.2. Did the ancient Greeks speak Greek?

Yes, the ancient Greeks did indeed speak Greek. The ancient Greek civilization, which emerged around 800 BC and lasted until about 600 AD, is well-known for its significant contributions to literature, philosophy, science, art, and politics. The primary language spoken by the ancient Greeks was Ancient Greek, which is the ancestor of the modern Greek language. Ancient Greek had various dialects, including Attic, Ionic, Doric, Aeolic, and Koine, each with its own distinct characteristics and variations.

The ancient Greeks used their language not only for everyday communication but also for the creation of a rich body of written works, including epic poems, philosophical texts, dramas, historical accounts, and more. These texts provide valuable insights into the culture, beliefs, and societal structure of ancient Greek civilization.

Factual correctness	Textual quality	Independence from baseline
Y	Y	N/A

EL

3.2. Οι αρχαίοι Έλληνες μιλούσαν ελληνικά;

Ναι, οι αρχαίοι Έλληνες μιλούσαν την αρχαία ελληνική γλώσσα. Η αρχαία ελληνική γλώσσα αποτελεί τον πρόγονο της σημερινής ελληνικής γλώσσας και υπήρξε η κύρια γλώσσα που χρησιμοποιούσαν οι αρχαίοι Έλληνες κατά την περίοδο της αρχαιότητας.

Η αρχαία ελληνική γλώσσα είχε *διάφορους* διαλέκτους, όπως *ο αττικός, ο ιωνικός, ο δωρικός, ο αιολικός και ο κοινή, καθένας με τα δικά του* χαρακτηριστικά και παραλλαγές.

Οι αρχαίοι Έλληνες χρησιμοποιούσαν αυτήν τη γλώσσα όχι μόνο για την επικοινωνία τους καθημερινά, αλλά και για τη δημιουργία ενός πλούσιου σώματος γραπτών έργων, περιλαμβανομένων επικών ποιημάτων, φιλοσοφικών κειμένων, δραμάτων, ιστορικών αφηγήσεων και πολλών άλλων. Αυτά τα κείμενα παρέχουν σημαντικές πληροφορίες για τον πολιτισμό, τις πεποιθήσεις και την κοινωνική δομή της αρχαίας ελληνικής κοινωνίας.

Factual correctness	Textual quality	Independence from baseline
Y	N	N

EN

3.3. Why do French people eat snails?

Eating snails, known as "escargot" in French cuisine, is a practice that dates back to ancient times. The tradition of consuming snails can be traced to the Roman era, where snails were considered a delicacy and a symbol of high social status.

Over time, the culinary use of snails evolved, and they became a popular dish in various cultures, including French cuisine. The French have refined the preparation of snails, often serving them with garlic, butter, and parsley to enhance the flavor. The appeal of eating snails lies in their unique taste and texture, which some people find enjoyable. Additionally, snails are a good source of protein and are low in fat, making them a relatively healthy food option.

In modern times, eating snails is not limited to the French; it's enjoyed in many countries around the world, particularly those with a rich culinary tradition. It's important to note that cultural preferences and culinary practices can vary widely, and what is considered a delicacy in one culture may not be in another.

Factual correctness	Textual quality	Independence from baseline
Y	Y	N/A

FR

3.3. Pourquoi les Français mangent-ils des escargots ?

Manger des escargots, *appelés "escargot"* en cuisine française, est une pratique culinaire qui remonte à l'Antiquité. La tradition de consommer des escargots remonte à l'époque romaine, où ils étaient considérés comme un mets délicat et un symbole de statut social élevé.

Au fil du temps, l'utilisation culinaire des escargots a évolué, et ils sont devenus un plat populaire dans diverses cultures, y compris la cuisine française. Les Français ont affiné la préparation des escargots, les servant souvent avec de l'ail, du beurre et du persil pour rehausser la saveur.

L'attrait de manger des escargots réside dans leur goût et leur texture uniques, que certaines personnes trouvent agréables. De plus, les escargots sont une bonne source de protéines et sont faibles en gras, ce qui en fait une option alimentaire relativement saine. De nos jours, manger des escargots n'est pas réservé aux seuls Français ; *c'est apprécié* dans de nombreux pays à travers le monde, en particulier ceux ayant une riche tradition culinaire. Il est important de noter que les préférences culturelles et les pratiques culinaires peuvent varier largement, et ce qui est considéré comme un mets délicat dans une culture peut ne pas l'être dans une autre.

Factual correctness	Textual quality	Independence from baseline
Y	N	N

EN

3.4. Why does Israel participate in Eurovision?

Israel participates in the Eurovision Song Contest because the Eurovision Song Contest is open to members of the European Broadcasting Union (EBU), and Israel is a member of the EBU. The Eurovision Song Contest is an annual international song competition, first held in 1956, where participating countries submit original songs to be performed on live television and then vote for the songs of other countries to determine the winner.

While Israel is not geographically located in Europe, its inclusion in the Eurovision Song Contest is based on its membership in the EBU, which is an organization of broadcasting organizations from different countries. The EBU extends its membership beyond Europe to include countries in the Mediterranean, North Africa, and the Middle East. Israel has been participating in the Eurovision Song Contest since 1973 and has won the competition multiple times.

The Eurovision Song Contest aims to bring together countries through music and cultural exchange, and its eligibility criteria are based on EBU membership rather than strict geographic or political boundaries.

Factual correctness	Textual quality	Independence from baseline
Y	Y	N/A

HE

3.4. באירוויזיון? משתתפת ישראל למה.

נציגים השולחות שונות מדינות מבין הזוכה את לקבץ שנועדה בינלאומית שירה תחרות הוא האירוויזיון השידור באיגוד לחברות ומיועדת, 1956 בשנת לראשונה נוצרה התחרות בטלוויזיה. חי* לביצוע בתחרות. להשתתף רשאית ולכן באיגוד, חברה היא ישראל. (EBU) האירופי השידור, באיגוד חברותה על מבוססת השתתפותה באירופה, גיאוגרפית ממקמת לא שישראל למרות פוליטיים*. או קפדניים גיאוגרפיים גבולות על ולא על מבוססים להשתתפות הקריטריונים ותרבות. מוזיקה דרך מדינות לאחד היא האירוויזיון מטרת כולל— התיכון ובמזרח אפריקה בצפון מדינות לכן, גיאוגרפי. מיקום על דווקא ולא באיגוד חברות באיגוד. חברות הן עוד כל להשתתף, יכולות— ישראל חי, בשידור משודרת התחרות פעמים. מספר זכתה ואף ב-1973 בתחרות לראשונה השתתפה ישראל לייצוגן. מקוריים שירים שולחות אשר מגוונות, מדינות משתתפות ובה

Factual correctness	Textual quality	Independence from baseline
Y	Y	Y

EN

3.5. Is it hard to work for a Japanese company?

Working in a Japanese company can be both rewarding and challenging, and the experience can vary based on factors such as the company culture, your role, and your personal adaptability. Here are some aspects to consider:

1. Corporate Culture: Japanese companies often have a strong emphasis on teamwork, harmony, and a hierarchical structure. Decision-making may be more consensus-based, and there is a focus on maintaining good relationships within the team.
2. Hierarchy and Respect: Respect for authority and seniority is crucial in Japanese corporate culture. There is a clear hierarchy, and employees are expected to show deference to their superiors. This can influence communication styles and decision-making processes.
3. Long Working Hours: In some Japanese companies, long working hours are common, and the concept of "salaryman" culture, where employees work extensive overtime, is prevalent. However, there have been efforts to address this issue, and some companies are adopting more flexible working arrangements.

4. Job Stability: Job security is often high in Japanese companies, and there is a traditional expectation of long-term employment. Loyalty to the company is valued, and employees may stay with one company for their entire career.
5. Language Barrier: Depending on the company and your role, language can be a potential challenge. While many international companies use English in the workplace, some traditional Japanese companies may conduct business primarily in Japanese.
6. Adaptability: Being adaptable and open to learning about and respecting Japanese customs and business etiquette is essential. Understanding and embracing the local culture can positively influence your experience.
7. Training and Development: Japanese companies often invest in training and development programs for employees. Continuous learning and improvement are valued, and employees may have opportunities to attend various workshops and training sessions.

Overall, the experience of working in a Japanese company can be enriching, providing insight into a unique corporate culture. However, the adjustment process might take time, especially if you are not familiar with the cultural nuances and business practices. Building strong relationships with colleagues, being open to learning, and showing respect for cultural differences can contribute to a successful experience in a Japanese workplace.

Factual correctness	Textual quality	Independence from baseline
Y	Y	N/A

JA

3.5. 日本の会社で働くのは大変ですか？

日本の会社で働くことは、やりがいがありつつも、課題があることがあります。経験は企業文化、あなたの役割、そして個人の適応力によって異なります。以下は考慮すべきいくつかの側面です：

1. 企業文化：日本の企業はチームワーク、調和、階層構造に力を入れることが一般的です。意思決定は合意形成が重視され、チーム内の良好な関係の維持が求められます。

2. 階層と尊敬：日本の企業文化では権威と先輩への尊敬が非常に重要です。明確な階層構造があり、部下は上司に対して敬意を払うことが期待されます。これがコミュニケーションスタイルや意思決定プロセスに影響を与える可能性があります。

3. 長時間労働：一部の日本の企業では長時間労働が一般的で、「サラリーマン」文化（固定給与の労働者が多くの時間働く傾向）が広く行われています。ただし、これに対処するための取り組みも行われており、柔軟な労働スケジュールを導入している企業もあります。

4. 雇用の安定性：雇用の安定性は一般的に高く、長期雇用が期待される伝統的な考え方があります。企業への忠誠心が重視され、従業員は通常、一つの企業で長く働くことがあります。

5. 言語の壁：企業や役割によっては、言語が潜在的な課題となることがあります。多くの国際企業は職場で英語を使用していますが、伝統的な日本企業では主に日本語で業務が行われることがあります。

6. 適応力: 日本の習慣やビジネスエチケットに対して適応力があり、* 学びに開かれている * ことが重要です。地元の文化を理解し、受け入れることが経験に良い影響を与える可能性があります。

7. トレーニングと開発: 日本の企業はしばしば従業員のトレーニングと開発プログラム * 投資しています *。継続的な学習と向上心が評価され、従業員は様々なワークショップやトレーニングセッションに参加する機会があります。

全体として、日本の企業で働く経験は豊かで、ユニークな企業文化の理解を提供する可能性があります。ただし、適応プロセスは時間がかかるかもしれません、特に文化的な * 微妙さ * やビジネスプラクティスに不慣れな場合 *。* 同僚との強い関係の構築、学びに開かれる姿勢、文化的な違いに対する尊重を示すことは、日本の職場 * で *。

Factual correctness	Textual quality	Independence from baseline
Y	N	N

EN

3.6. What is the weather like in Russia?

Russia is an enormous country with diverse climates due to its vast size and geographical features. As a result, the weather can vary significantly depending on the region and the time of year. Here's a general overview:

1. European Russia: In the western part of the country, where cities like Moscow and St. Petersburg are located, the climate is predominantly continental with distinct seasons. Winters are cold, often with snow, while summers can be warm.

2. Siberia: Eastern Russia, particularly Siberia, experiences an extreme continental climate. Winters are long and bitterly cold, with temperatures often dropping well below freezing. Summers can be relatively warm, but they are short.

3. Far East: The Russian Far East, including Vladivostok, has a humid continental climate. Winters are cold, and summers are relatively short with moderate temperatures.

4. Northern Russia: The northern regions, such as Murmansk, experience a subarctic or Arctic climate, characterized by very cold temperatures, especially in winter.

5. Southern Russia: Regions in the south, like Sochi on the Black Sea coast, have a milder climate with relatively mild winters and warm summers.

It's important to note that Russia's climate diversity means that weather conditions can be drastically different from one region to another. For detailed and up-to-date information on the current weather in a specific location, it's recommended to check local weather forecasts or reliable online weather services.

Factual correctness	Textual quality	Independence from baseline
Y	Y	N/A

RU

3.6. Какая погода в России?

Погода в России разнообразна и зависит от географических и климатических особенностей различных регионов. Вот общий обзор:

1. Европейская часть России: В западной части, где расположены города, такие как Москва и Санкт-Петербург, климат в основном континентальный с четко выраженными сезонами. Зимы холодные, часто с снегопадами, а лета могут быть теплыми.
2. Сибирь: В восточной части России, особенно в Сибири, преобладает суровый континентальный климат. Зимы длинные и морозные, с температурами часто значительно ниже нуля. Лета могут быть относительно теплыми, но короткими.
3. Дальний Восток: Дальний Восток России, включая Владивосток, имеет влажный континентальный климат. Зимы холодные, а лета относительно короткие с умеренными температурами.
4. Север России: Северные регионы, такие как Мурманск, имеют субарктический или арктический климат с очень холодными температурами, особенно зимой.
5. Юг России: Регионы на юге, такие как Сочи на побережье Черного моря, имеют более мягкий климат с относительно теплыми зимами и жаркими летами.

Важно отметить, что климатическое разнообразие России означает, что условия могут существенно различаться от региона к региону. Для получения подробной и актуальной информации о погоде в конкретном регионе рекомендуется проверять местные прогнозы или использовать надежные онлайн-сервисы погоды.

Factual correctness	Textual quality	Independence from baseline
Y	Y	N

B. Full Answers and Scores (Question 3): GPT-4.5

Language mistakes are denoted between asterisk (*) symbols.

3. Geography-/Culture-Sensitive Questions
EN
3.1. Is Bulgarian the oldest Slavic language?
Bulgarian is not the oldest Slavic language, but it is closely related to Old Church Slavonic, which is considered the first literary Slavic language. Old Church Slavonic (also known as Old Bulgarian) emerged in the 9th century, when Saints Cyril and Methodius created the Glagolitic alphabet and translated Christian religious texts for the Slavic-speaking peoples of Great Moravia. Their disciples later brought these texts to Bulgaria, where the Cyrillic alphabet was developed in the Preslav Literary School, forming the foundation of Bulgarian literary tradition. Modern Bulgarian evolved significantly from Old Church Slavonic over the centuries, undergoing considerable linguistic changes, including loss of case endings and development of a definite article—distinct from most other Slavic languages.

Thus, while Modern Bulgarian is closely connected historically to Old Church Slavonic, it itself is not the oldest Slavic language. Rather, Old Church Slavonic is recognized as the earliest documented Slavic literary language, with Bulgarian being its direct descendant.

Factual correctness	Textual quality	Independence from baseline
Y	Y	N/A

BG

3.1. Българският най-старият славянски език ли е?

Българският не е най-старият славянски език, защото всички славянски езици произлизат от общ праславянски език, който не е запазен в писмен вид.

Все пак българският е първият славянски език, документиран писмено и използван като литературен и богослужебен език, благодарение на старобългарския (старославянския), възникнал през IX век в Първото българско царство.

Старобългарският (наричан и старославянски или църковнославянски) е първият писмено засвидетелстван славянски език и основа за развитието на множество други славянски езици и култури. В този смисъл българският език играе особена историческа роля, но строго погледнато, най-старият славянски език е праславянският, от който произлизат всички съвременни славянски езици.

Factual correctness	Textual quality	Independence from baseline
Y	Y	Y

EN

3.2. Did the ancient Greeks speak Greek?

Yes, the ancient Greeks spoke Greek, specifically ancient Greek, which is distinct from the modern Greek spoken today. Ancient Greek had several dialects, including Attic, Ionic, Doric, Aeolic, and later Koine Greek, which became widespread after the conquests of Alexander the Great. Koine Greek was also the language of the New Testament and served as a common language across the eastern Mediterranean for several centuries.

Factual correctness	Textual quality	Independence from baseline
Y	Y	N/A

EL

3.2. Οι αρχαίοι Έλληνες μιλούσαν ελληνικά;

Ναί, οι αρχαίοι Έλληνες έλληνιστί έλάλουν.

Factual correctness	Textual quality	Independence from baseline
Y	Y	Y

EN

3.3. Why do French people eat snails?

Eating snails, known as escargots, is a cultural tradition in France dating back to Roman times. Here's why they became a part of French cuisine:

1. Historical Origins:

Snails have been consumed since ancient Roman times, with archaeological evidence showing their popularity as a source of food.

The Romans introduced snail-eating traditions throughout Gaul (present-day France).

2. Culinary Tradition:

In medieval times, snails became associated with religious fasting periods, especially Lent, because they were not considered meat.

Snails eventually evolved into a delicacy, with chefs developing specialized recipes featuring garlic, butter, parsley, and wine, particularly in the Burgundy region.

3. Regional Specialties:

Burgundy (Bourgogne): Most famous for their escargots de Bourgogne—snails prepared with garlic-parsley butter.

Alsace and Provence also have their own regional snail recipes.

4. Gastronomic Status:

Snails have come to symbolize French culinary refinement and sophistication. Eating escargots today is considered a gourmet experience, often associated with special occasions.

5. Nutritional Benefits:

Snails are high in protein, low in fat, and contain beneficial nutrients like iron, magnesium, and selenium.

Today, the French don't eat snails daily; rather, they are enjoyed as a traditional delicacy, particularly in restaurants or festive meals, reflecting France's rich culinary heritage.

Factual correctness	Textual quality	Independence from baseline
Y	Y	N/A

FR

3.3. Pourquoi les Français mangent-ils des escargots ?

Les Français mangent des escargots essentiellement pour des raisons culturelles et gastronomiques :

Tradition culinaire ancienne :

La consommation d'escargots remonte à l'époque romaine et s'est développée particulièrement au Moyen Âge en France, où les escargots étaient abondants et facilement accessibles, notamment en Bourgogne.

Appréciation du goût :

Les escargots sont appréciés pour leur texture tendre, légèrement ferme, ainsi que pour leur capacité à absorber les saveurs d'accompagnements tels que l'ail, le beurre persillé et les fines herbes.

Produit gastronomique raffiné :

Considérés comme un mets fin, les escargots sont souvent associés aux repas festifs ou familiaux, comme Noël ou les fêtes de fin d'année.

Aspect identitaire et culturel :

Les escargots sont aujourd'hui emblématiques de la cuisine française à travers le monde. Ils participent au patrimoine culinaire français et à l'image gastronomique internationale du pays.

Même si tous les Français ne consomment pas régulièrement des escargots, ce plat reste une spécialité très appréciée et un symbole fort de l'art culinaire français.

Factual correctness	Textual quality	Independence from baseline
Y	Y	Y
EN		
3.4. Why does Israel participate in Eurovision?		
Israel participates in the Eurovision Song Contest because its national public broadcaster, the Israeli Public Broadcasting Corporation (IPBC/Kan), is an active member of the European Broadcasting Union (EBU). Membership in the EBU grants eligibility to participate in Eurovision, irrespective of a country’s geographical location. Israel’s involvement dates back to 1973, and the country has since secured four victories in the competition. {source link} The EBU encompasses broadcasters from across Europe and the Middle East, making the contest accessible to a diverse range of countries. Israel’s long-standing membership and active participation underscore the inclusive nature of Eurovision, which aims to unite audiences through music. {source link}		
Factual correctness	Textual quality	Independence from baseline
Y	Y	N/A
HE		
3.4. באירוויזיון? משתתפת ישראל למה.		
האירופי השידור באיגוד חברה הייתה הישראלית השידור שרשות משום באירוויזיון משתתפת ישראל תלויה אינה ההשתתפות בתחרות. להשתתפות התנאי היא באיגוד והחברות ה-50, שנות מאז (EBU) נוספות מדינות כמו ישראל, ולכן האירופי, השידור בארגון בחברות אלא באירופה, גאוגרפי במיקום האירוויזיון לתחרות הצטרפה ישראל קבוע. באופן משתתפת אוסטרליה), (למשל ליבשת מחוץ בשל להשתתף וממשיכה ו-2018) 1978, 1979, 1998 פעמים ארבע זכתה, 1973 בשנת לראשונה במדינה. התחרות של הרבה הפופולריות		
Factual correctness	Textual quality	Independence from baseline
Y	Y	Y
EN		
3.5. Is it hard to work for a Japanese company?		
Working for a Japanese company can be challenging for many people, but the difficulty depends greatly on the individual’s cultural background, work style, and expectations. Here are some common factors that can make it challenging: 1. Work Culture and Hours Japanese companies traditionally expect long working hours, including overtime. Commitment, loyalty, and company dedication are strongly emphasized. Taking frequent vacations or leaving work precisely at the end of your scheduled hours might be viewed negatively. 2. Hierarchy and Communication Hierarchical structures are prevalent, with clearly defined roles and seniority. Decision-making can be slower, with multiple layers of approvals required. Indirect communication is common; clear criticism or direct feedback may be avoided, making it harder for foreigners accustomed to straightforward discussions. 3. Language Barrier		

Japanese proficiency is often crucial, even if the company’s official business language is English.

Many internal documents, meetings, and day-to-day communications will likely occur in Japanese.

4. Group Harmony (Wa 和)

Harmony within the group is prioritized, sometimes above individual needs. Expressing disagreement openly can be considered disruptive, so employees often need to navigate social cues carefully.

5. Expectations and Pressure

High expectations regarding quality, punctuality, attention to detail, and consistency. Performance pressure can be significant, with regular evaluations and feedback.

Potential Benefits and Positive Aspects

On the other hand, there are positive aspects to working in a Japanese company: Stability: Many Japanese companies provide job security, stable income, and comprehensive benefits.

Professional Growth: Exposure to disciplined, meticulous business practices that emphasize quality and efficiency.

Skill Development: Opportunity to gain proficiency in Japanese and develop inter-cultural competencies.

Team Orientation: Emphasis on teamwork and collective achievement, which can create a strong sense of belonging and cooperation.

Recommendations:

Develop at least intermediate Japanese language proficiency.

Familiarize yourself with Japanese business etiquette and social customs.

Be open-minded, patient, and willing to adapt to a culturally different working environment.

In short, working at a Japanese company can be rewarding, but it requires patience, flexibility, and a willingness to adapt to cultural differences.

Factual correctness	Textual quality	Independence from baseline
Y	Y	N/A

JA

3.5. 日本の会社で働くのは大変ですか？

日本の会社で働くのは一般的に大変だと言われます。業界や会社、職種によっても違いますが、多くの場合以下のような特徴が挙げられます。

大変と感じる主な理由：

労働時間が長い

日本の会社では残業が比較的多く、「定時退社」が難しい職場もまだまだあります。

上下関係やマナーが厳しい

上司や先輩に対して丁寧な敬語を使い、組織の中での立場を強く意識しなければなりません。

有給休暇が取りにくい

日本では有給休暇を自由に取づらい雰囲気が残っている企業も多くあります。

集団行動が求められる

チームや組織内の協調性を重視するため、個人主義的な働き方が難しい場合もあります。

暗黙のルールが多い

明文化されていないルールやマナーが多く、外国人や新入社員にとっては馴染むまで苦労することがあります。

一方で、改善の傾向も：

最近では働き方改革やワークライフバランスを重視する企業も増え、残業規制やテレワーク、柔軟な働き方を取り入れる企業も出てきています。

ITや外資系、ベンチャー企業などでは比較的自由的な働き方が可能な場合もあります。

結論：

日本の会社で働くことは、文化や労働慣習の違いから特に外国人にとっては大変だと感じる人が多いですが、業界・企業によって差があります。職場選びを慎重に行えば、ストレスを軽減することも可能です。

Factual correctness	Textual quality	Independence from baseline
Y	Y	Y

EN

3.6. What is the weather like in Russia?

Russia is the largest country in the world, spanning multiple climate zones, so weather conditions can vary significantly across different regions. Here's an overview of current weather conditions and forecasts for several key cities:

В момента е 10° ·Ясно

Москва, Русия

Днес 14° 3° Слънчево

петък 16° 1° Ветровито

събота 6° -3° По-студено

неделя 0° -3° Сняг

понеделник 2° -5° Няколко превалявания от сняг

вторник 4° -4° Слънчево

сряда 3° -3° Облачно

В момента е 9° ·Ясно

Сейнт Питърсбург, Русия

Днес 12° 3° Облаци и слънце

петък 8° -1° Ветровито

събота 2° -3° Малко снеговалеж

неделя 3° -2° Слънчево

понеделник 5° -1° Няколко превалявания от сняг

вторник 7° 0° Облаци и слънце

сряда 6° -2° Няколко превалявания, предимно в по-късните часове

В момента е 6° ·Облачно

Новосибирск, Русия

Днес 9° 5° Ветровито

петък 17° 1° По-топло

събота 6° 3° Значително по-студено
неделя 12° 2° Облачно
понеделник 16° 5° Сгъстяваща се облачност
вторник 15° 3° Облачно
сряда 14° 2° Частично слънчево
В момента е 3° ·Дъжд
Екатеринбург, Русия
Днес 7° 0° Дъжд
петък 4° 1° Ниски облаци
събота 9° -1° Възможно преваляване
неделя 10° 1° Облаци и слънце
понеделник 5° 0° Слаб валеж от дъжд
вторник 8° 1° Предимно облачно
сряда 8° 1° Слаб валеж от дъжд

As seen, western cities like Moscow and Saint Petersburg are experiencing relatively mild temperatures with a mix of sun and clouds, while eastern cities like Novosibirsk and Yekaterinburg are cooler with chances of precipitation. It’s advisable to check the specific forecasts for your destination, as conditions can vary widely across Russia.

Factual correctness	Textual quality	Independence from baseline
N	Y	N/A

RU

3.6. Какая погода в России?

Россия — страна с обширной территорией, поэтому погодные условия значительно различаются в разных регионах. Ниже представлены текущие погодные условия и прогнозы для некоторых крупных городов России:

Москва:
В настоящее время 10° ·Ясно
Москва, Россия
Сегодня 14° 3° Солнечно
пятница 16° 1° Ветрено
суббота 6° -3° Холоднее
воскресенье 0° -3° Снег
понеделник 2° -5° Кратковременный снегопад
вторник 4° -4° Солнечно
среда 3° -3° Облачно
Санкт-Петербург:
В настоящее время 9° ·Ясно
Санкт-Петербург, Россия
Сегодня 12° 3° Переменная облачность
пятница 8° -1° Ветрено
суббота 2° -3° Небольшой снег
воскресенье 3° -2° Солнечно
понеделник 5° -1° Кратковременный снегопад
вторник 7° 0° Переменная облачность

среда 6° -2° Кратковременные ливни, преимущественно в позднее время
Новосибирск:
В настоящее время 6° ·Облачно
Новосибирск, Россия
Сегодня 9° 5° Ветрено
пятница 17° 1° Теплее
суббота 6° 3° Намного холоднее
воскресенье 12° 2° Облачно
понедельник 16° 5° Растущая облачность
вторник 15° 3° Облачно
среда 14° 2° Небольшая облачность
Для получения более точной информации о погоде в вашем регионе
рекомендуется воспользоваться местными метеорологическими службами
или онлайн-сервисами.

Factual correctness	Textual quality	Independence from baseline
Y	Y	N

C. Statistical Analysis

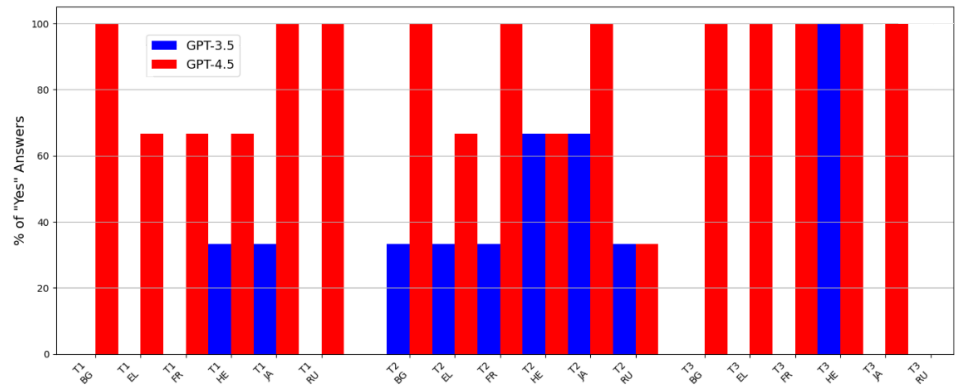


Figure 7
Percentage of 'Yes' scores for 'Independence from Baseline' per question type and per language

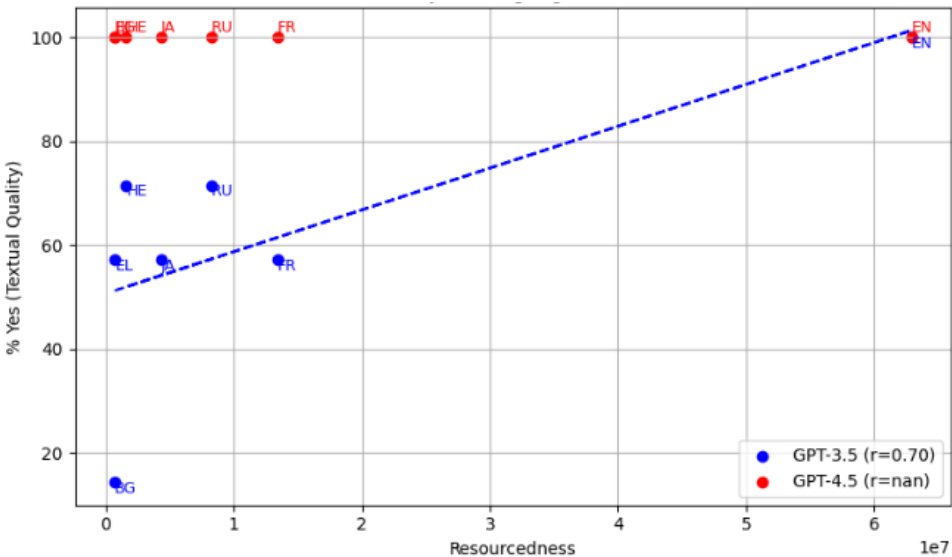


Figure 8
Resourcedness vs ‘Textual’ Quality
* The number of Wikipedia pages in the language is taken as a proxy for resourcedness.

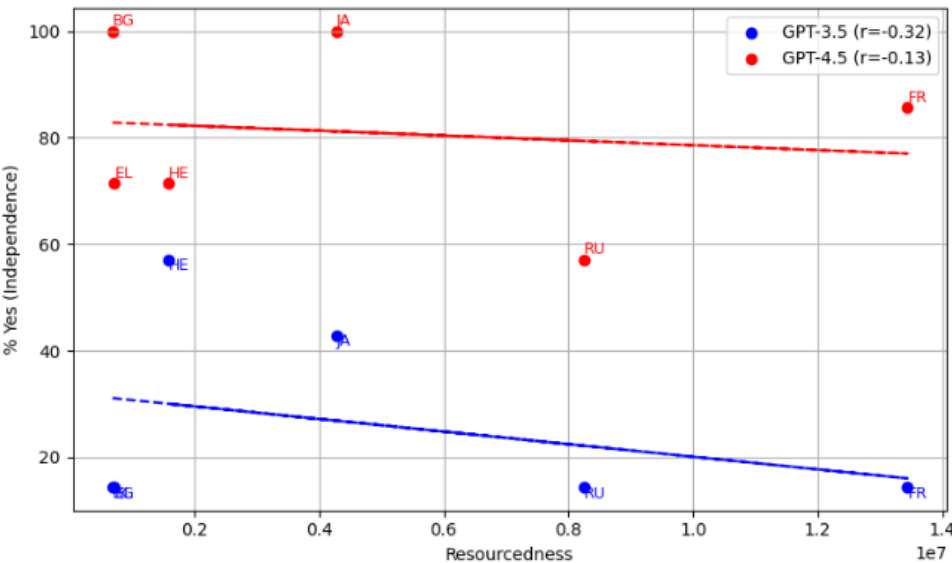


Figure 9
Resourcedness vs ‘Independence from Baseline’
* The number of Wikipedia pages in the language is taken as a proxy for resourcedness.