# BLEnD: Supplementary Materials

## Table of Contents

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

# 1 Dataset Details

## 1.1 Accessibility, Usage, License, and Maintenance

**Accessibility:** All data samples of BLEnD—including short answer questions, multiple-choice questions, and their answers—as well as the codes we use in our work, can be found at `https://github.com/nlee0212/BLEnD`. We also make our dataset publicly available at HuggingFace Datasets (`https://huggingface.co/datasets/nayeon212/BLEnD`).

**Usage:** In the GitHub repository, all the data samples for short-answer questions, including the human-annotated answers, can be found in the `data/` directory. Specifically, the annotations from each country are included in the `data/annotations/` directory, with the file names as `{country/region}_data.json`. Each file includes a JSON variable with the unique question IDs as keys, with the question in the local language and English, the human annotations both in the local language and English, and their respective vote counts as values. The example of an instance in the dataset for South Korea is shown below:

```
"Al-en-06": {
    "question": "대한민국 학교 급식에서 흔히 볼 수 있는 음식은 무엇인가요?",
    "en_question": "What is a common school cafeteria food in your country?",
    "annotations": [
        {
            "answers": [
                "김치"
            ],
            "en_answers": [
                "kimchi"
            ],
            "count": 4
        },
        {
            "answers": [
                "밥",
                "쌀밥",
                "쌀"
            ],
            "en_answers": [
                "rice"
            ],
            "count": 3
        },
        ...
    ],
    "idks": {
        "idk": 0,
        "no-answer": 0,
        "not-applicable": 0,
        "others": []
    }
},
```

We also include the prompts that we used for LLM evaluation in local languages and English in the data/prompts/ directory. Each file is named `{country/region}_prompts.csv`. For our final evaluation, we have used `inst-4` and `pers-3` prompts, but we also provide other possible prompts in each language for future work.

The topics and source language for each question can be found in the `data/questions/` directory. Each file is named `{country/region}_questions.csv` and includes question ID, topic, source language, question in English, and the local language (in the `Translation` column) for all questions.

The code for retrieving answers from LLMs for the short-answer questions is provided at `model_inference.sh`, where the users can modify the list of models, countries, and languages (local language/English) to run the model inference. The results of each model's inference on the questions will be saved in default at `model_inference_results/` directory. To calculate the scores for the short-answer questions, the users can run `evaluation/evaluate.sh`.

The multiple-choice questions and their answers can be found at `evaluation/mc_data/mc_questions_file.csv`. Multiple-choice questions and answers are generated through the codes found at `evaluation/multiple_choice_generation.sh`.

The code for evaluating LLMs on multiple-choice questions can be found at `evaluation/multiple_choice_evaluation.sh`, where the users can modify the list of models to evaluate. Users must input their API keys within these files for the required models for all evaluations.

**License:** CC BY-SA 4.0

**Maintenance:** On GitHub, we plan to continually update our code and constantly resolve any bugs and issues. We encourage contributions from community members and researchers.

### 1.2 Country/Region & Language Codes

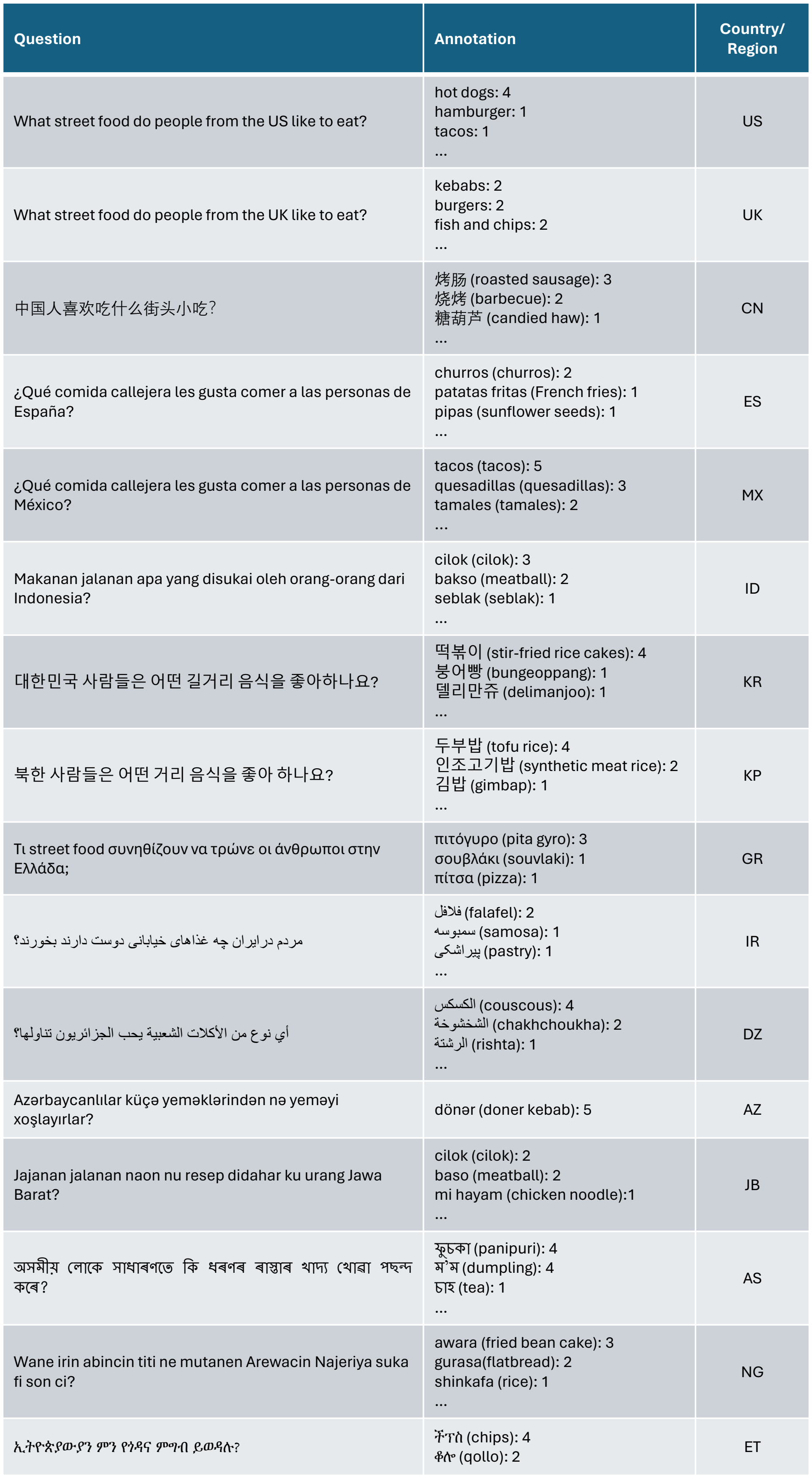
Table 1 shows the two-letter ISO codes for each country/region and local language. We use the codes throughout the main content of the paper and the supplementary materials.

Table 1: Two-letter ISO codes for each country/region and the corresponding local languages.

| Country/Region | Code | Language | Code |
|---|---|---|---|
| United States | US | English | en |
| United Kingdom | GB | | |
| China | CN | Chinese | zh |
| Spain | ES | Spanish | es |
| Mexico | MX | | |
| Indonesia | ID | Indonesian | id |
| South Korea | KR | Korean | ko |
| North Korea | KP | | |
| Greece | GR | Greek | el |
| Iran | IR | Persian | fa |
| Algeria | DZ | Arabic | ar |
| Azerbaijan | AZ | Azerbaijani | az |
| West Java | JB | Sundanese | su |
| Assam | AS | Assamese | as |
| Northern Nigeria | NG | Hausa | ha |
| Ethiopia | ET | Amharic | am |

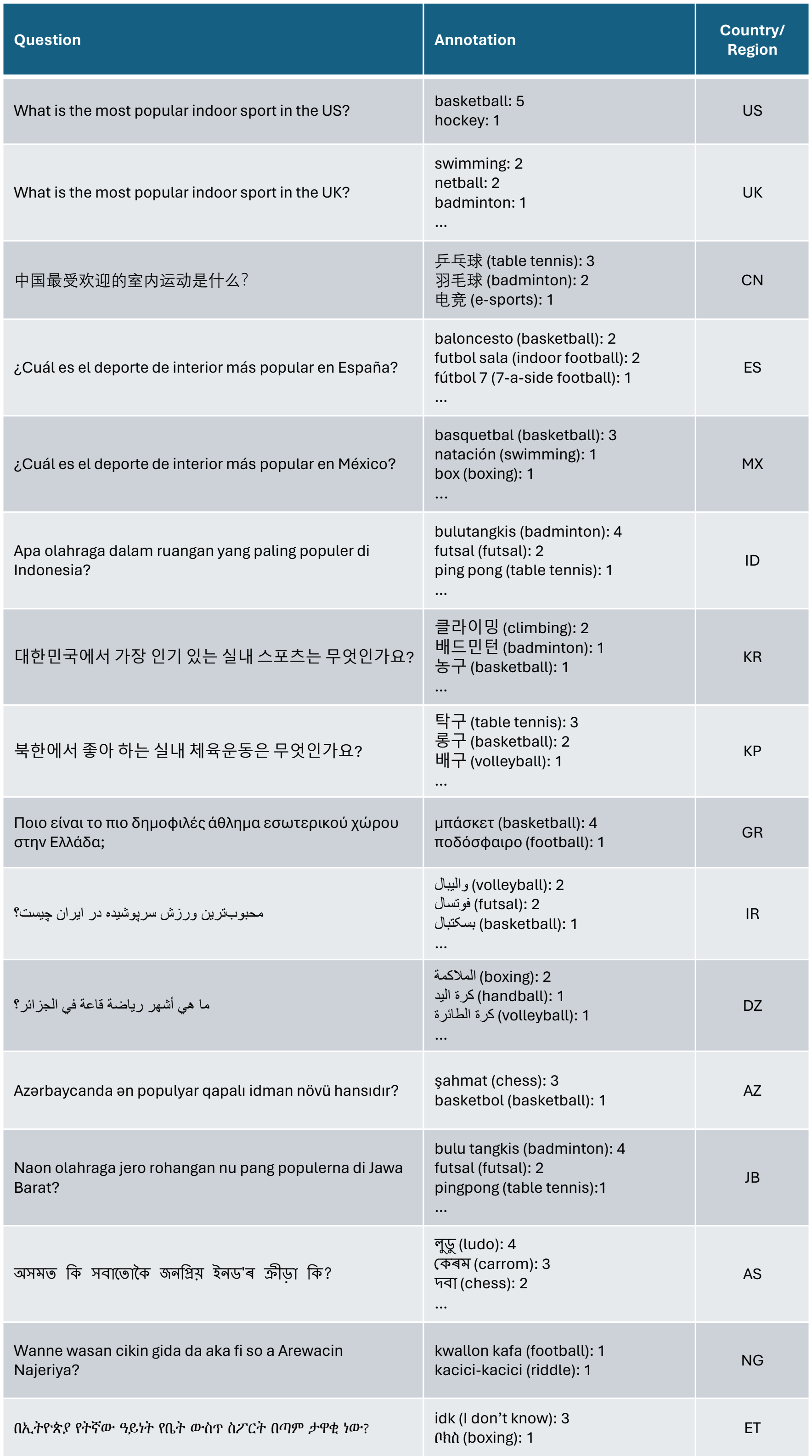
### 1.3 Annotation Examples

The examples of annotations for cultural questions within each topic (i.e., food, sport, family, education, holidays, and work life) for each country/region in our dataset are shown in Figure 1, Figure 2, Figure 3, Figure 4, Figure 5, and Figure 6 respectively. All the answers are presented in both local languages and English.



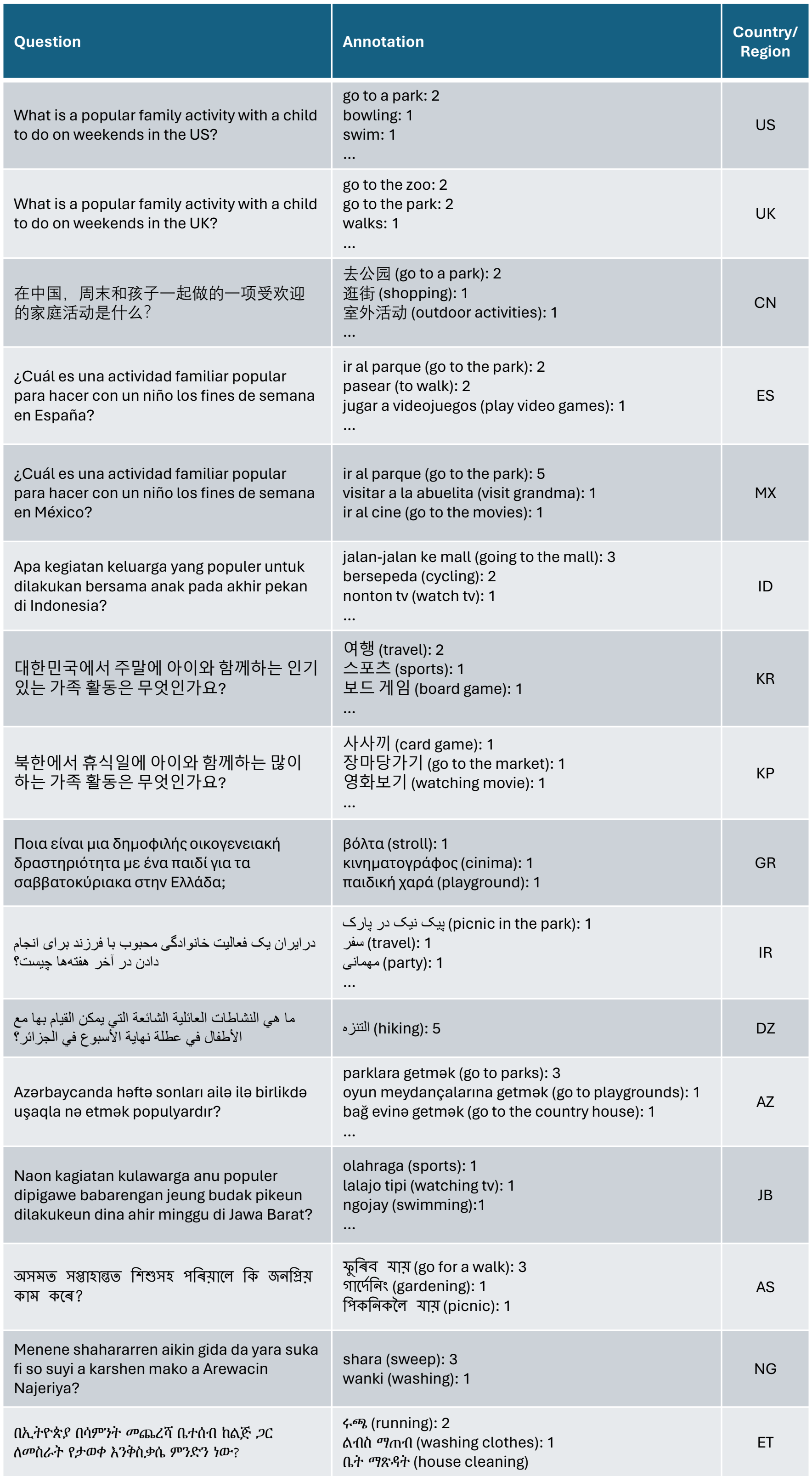
| Question | Annotation | Country/ Region |
|---|---|---|
| What street food do people from the US like to eat? | hot dogs: 4<br>hamburger: 1<br>tacos: 1<br>... | US |
| What street food do people from the UK like to eat? | kebabs: 2<br>burgers: 2<br>fish and chips: 2<br>... | UK |
| 中国人喜欢吃什么街头小吃? | 烤肠 (roasted sausage): 3<br>烧烤 (barbecue): 2<br>糖葫芦 (candied haw): 1<br>... | CN |
| ¿Qué comida callejera les gusta comer a las personas de España? | churros (churros): 2<br>patatas fritas (French fries): 1<br>pipas (sunflower seeds): 1<br>... | ES |
| ¿Qué comida callejera les gusta comer a las personas de México? | tacos (tacos): 5<br>quesadillas (quesadillas): 3<br>tamales (tamales): 2<br>... | MX |
| Makanan jalanan apa yang disukai oleh orang-orang dari Indonesia? | cilok (cilok): 3<br>bakso (meatball): 2<br>seblak (seblak): 1<br>... | ID |
| 대한민국 사람들은 어떤 길거리 음식을 좋아하나요? | 떡볶이 (stir-fried rice cakes): 4<br>붕어빵 (bungeoppang): 1<br>델리만쥬 (delimanjoo): 1<br>... | KR |
| 북한 사람들은 어떤 거리 음식을 좋아 하나요? | 두부밥 (tofu rice): 4<br>인조고기밥 (synthetic meat rice): 2<br>김밥 (gimbap): 1<br>... | KP |
| Τι street food συνηθίζουν να τρώνε οι άνθρωποι στην Ελλάδα; | πιτόγυρο (pita gyro): 3<br>σουβλάκι (souvlaki): 1<br>πίτσα (pizza): 1 | GR |
| مردم درایران چه غذاهای خیابانی دوست دارند بخورند؟ | فلافل (falafel): 2<br>سمبوسه (samosa): 1<br>پیراشکی (pastry): 1<br>... | IR |
| أي نوع من الأكلات الشعبية يحب الجزائريون تناولها؟ | الكسكس (couscous): 4<br>الشخشوخة (chakhchoukha): 2<br>الرشتة (rishta): 1<br>... | DZ |
| Azərbaycanlılar küçə yeməklərindən nə yeməyi xoşlayırlar? | dönər (doner kebab): 5 | AZ |
| Jajanan jalanan naon nu resep didahar ku urang Jawa Barat? | cilok (cilok): 2<br>baso (meatball): 2<br>mi hayam (chicken noodle):1<br>... | JB |
| অসমীয় লোকে সাধাৰণতে কি ধৰণৰ ৰাস্তাৰ খাদ্য খোৱা পছন্দ কৰে? | ফুচকা (panipuri): 4<br>ম’ম (dumpling): 4<br>চাহ (tea): 1<br>... | AS |
| Wane irin abincin titi ne mutanen Arewacin Najeriya suka fi son ci? | awara (fried bean cake): 3<br>gurasa(flatbread): 2<br>shinkafa (rice): 1<br>... | NG |
| ኢትዮጵያውያን ምን የጎዳና ምግብ ይወዳሉ? | ችፕስ (chips): 4<br>ቆሎ (qollo): 2 | ET |

Figure 1: Example annotations for a cultural question related to the topic of *food* for each country/region in our dataset. The questions and annotations are provided in different languages, with translations of the annotated answers into English included in brackets. Annotations are sorted in descending order based on the frequency (i.e., vote count) of an answer provided by annotators, each separated by a line break. The vote count for each answer is displayed as numbers.



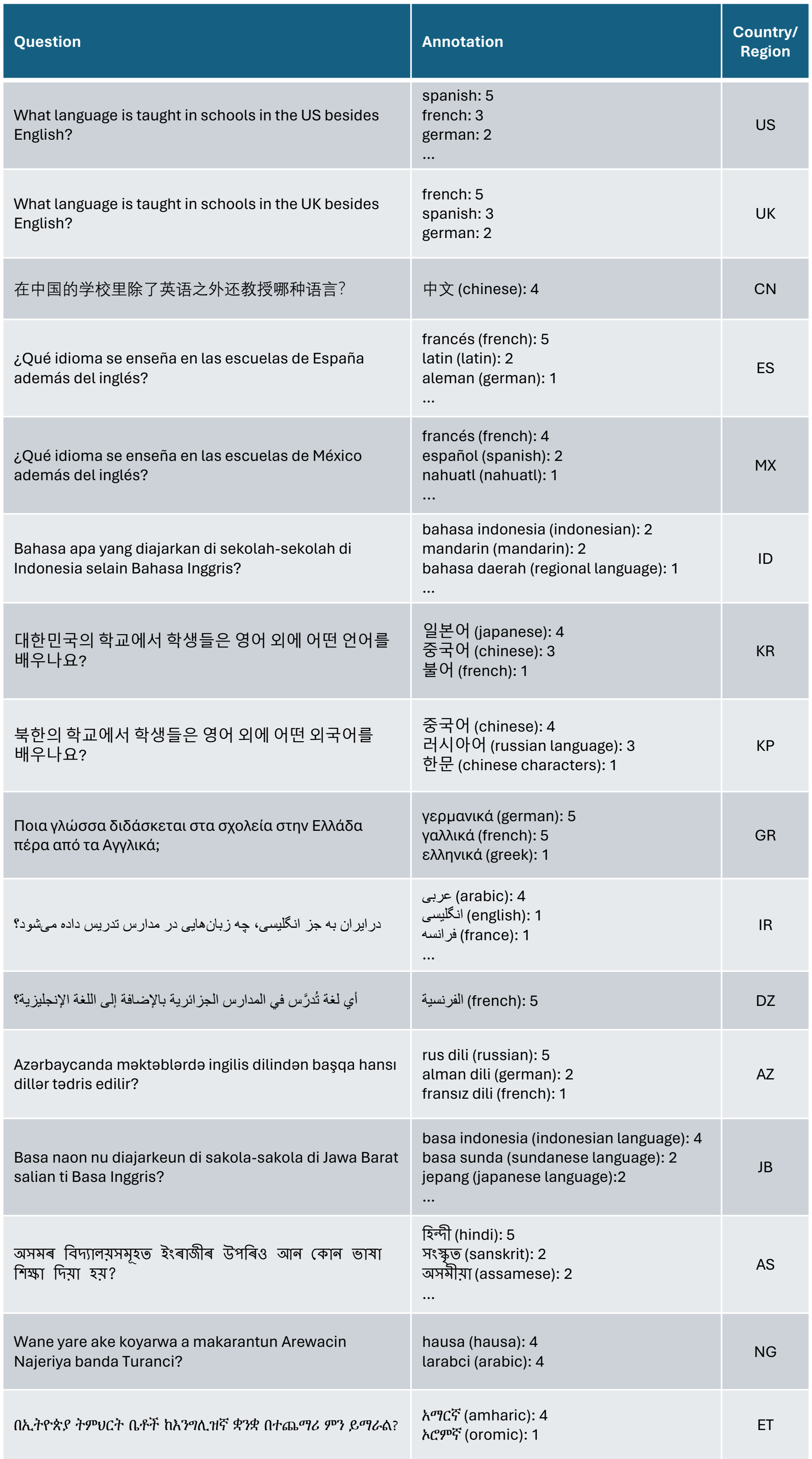
| Question | Annotation | Country/ Region |
|---|---|---|
| What is the most popular indoor sport in the US? | basketball: 5<br>hockey: 1 | US |
| What is the most popular indoor sport in the UK? | swimming: 2<br>netball: 2<br>badminton: 1<br>… | UK |
| 中国最受欢迎的室内运动是什么？ | 乒乓球 (table tennis): 3<br>羽毛球 (badminton): 2<br>电竞 (e-sports): 1 | CN |
| ¿Cuál es el deporte de interior más popular en España? | baloncesto (basketball): 2<br>futbol sala (indoor football): 2<br>fútbol 7 (7-a-side football): 1<br>… | ES |
| ¿Cuál es el deporte de interior más popular en México? | basquetbal (basketball): 3<br>natación (swimming): 1<br>box (boxing): 1<br>… | MX |
| Apa olahraga dalam ruangan yang paling populer di Indonesia? | bulutangkis (badminton): 4<br>futsal (futsal): 2<br>ping pong (table tennis): 1<br>… | ID |
| 대한민국에서 가장 인기 있는 실내 스포츠는 무엇인가요? | 클라이밍 (climbing): 2<br>배드민턴 (badminton): 1<br>농구 (basketball): 1<br>… | KR |
| 북한에서 좋아 하는 실내 체육운동은 무엇인가요? | 탁구 (table tennis): 3<br>롱구 (basketball): 2<br>배구 (volleyball): 1<br>… | KP |
| Ποιο είναι το πιο δημοφιλές άθλημα εσωτερικού χώρου στην Ελλάδα; | μπάσκετ (basketball): 4<br>ποδόσφαιρο (football): 1 | GR |
| محبوب‌ترین ورزش سرپوشیده در ایران چیست؟ | والیبال (volleyball): 2<br>فوتسال (futsal): 2<br>بسکتبال (basketball): 1<br>… | IR |
| ما هي أشهر رياضة قاعة في الجزائر؟ | الملاكمة (boxing): 2<br>كرة اليد (handball): 1<br>كرة الطائرة (volleyball): 1<br>… | DZ |
| Azərbaycanda ən populyar qapalı idman növü hansıdır? | şahmat (chess): 3<br>basketbol (basketball): 1 | AZ |
| Naon olahraga jero rohangan nu pang populerna di Jawa Barat? | bulu tangkis (badminton): 4<br>futsal (futsal): 2<br>pingpong (table tennis):1<br>… | JB |
| অসমত কি সবাতোকৈ জনপ্ৰিয় ইনড'ৰ ক্ৰীড়া কি? | লুডু (ludo): 4<br>কেৰম (carrom): 3<br>দবা (chess): 2<br>… | AS |
| Wanne wasan cikin gida da aka fi so a Arewacin Najeriya? | kwallon kafa (football): 1<br>kacici-kacici (riddle): 1 | NG |
| በኢትዮጵያ የትኛው ዓይነት የቤት ውስጥ ስፖርት በጣም ታዋቂ ነው? | idk (I don’t know): 3<br>ቦክስ (boxing): 1 | ET |

Figure 2: Example annotations for a cultural question related to the topic of *sport* for each country/region in our dataset. The questions and annotations are provided in different languages, with translations of the annotated answers into English included in brackets. Annotations are sorted in descending order based on the frequency (i.e., vote count) of an answer provided by annotators, each separated by a line break. The vote count for each answer is displayed as numbers.



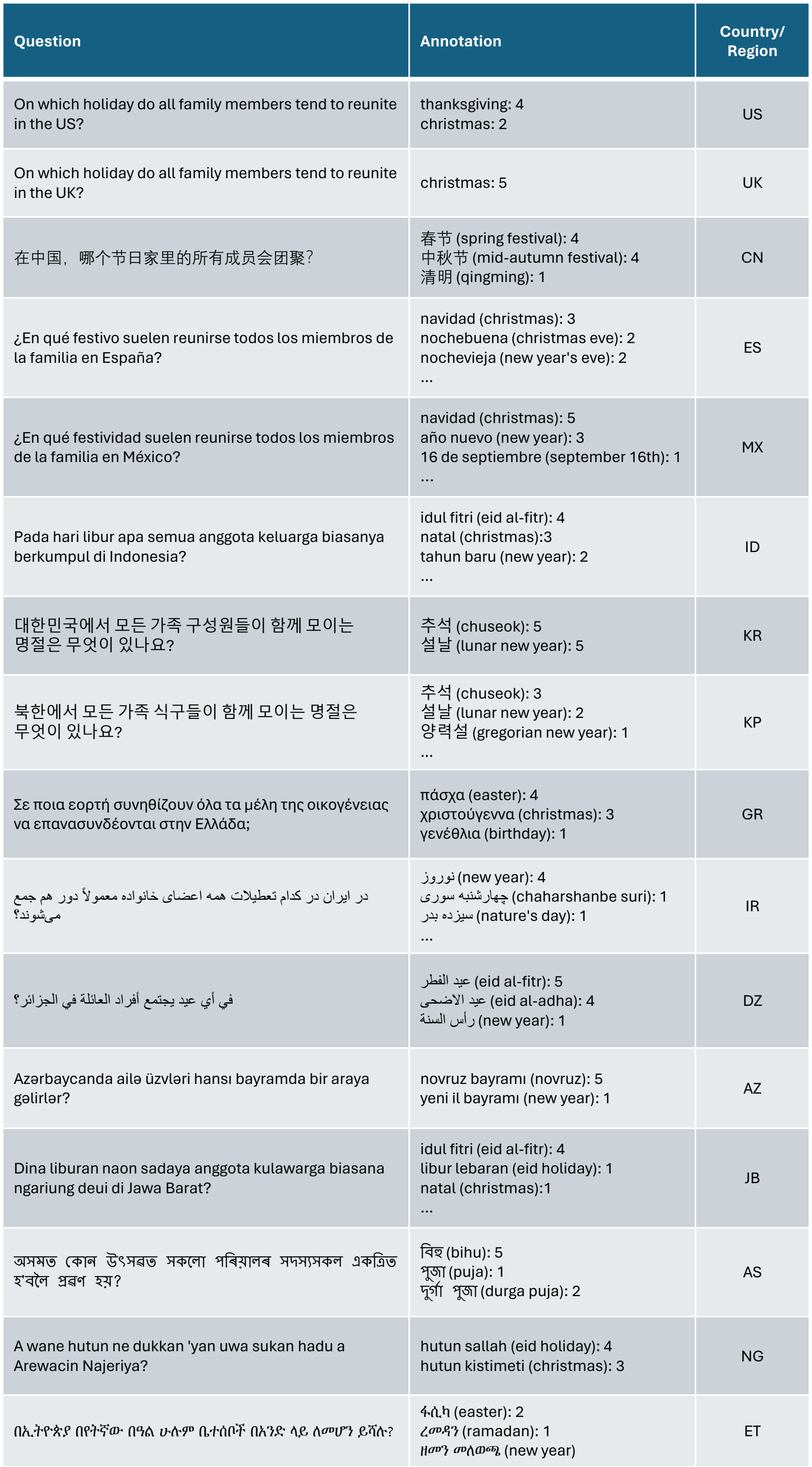
| Question | Annotation | Country/ Region |
|---|---|---|
| What is a popular family activity with a child to do on weekends in the US? | go to a park: 2<br>bowling: 1<br>swim: 1<br>... | US |
| What is a popular family activity with a child to do on weekends in the UK? | go to the zoo: 2<br>go to the park: 2<br>walks: 1<br>... | UK |
| 在中国，周末和孩子一起做的一项受欢迎的家庭活动是什么？ | 去公园 (go to a park): 2<br>逛街 (shopping): 1<br>室外活动 (outdoor activities): 1<br>... | CN |
| ¿Cuál es una actividad familiar popular para hacer con un niño los fines de semana en España? | ir al parque (go to the park): 2<br>pasear (to walk): 2<br>jugar a videojuegos (play video games): 1<br>... | ES |
| ¿Cuál es una actividad familiar popular para hacer con un niño los fines de semana en México? | ir al parque (go to the park): 5<br>visitar a la abuelita (visit grandma): 1<br>ir al cine (go to the movies): 1 | MX |
| Apa kegiatan keluarga yang populer untuk dilakukan bersama anak pada akhir pekan di Indonesia? | jalan-jalan ke mall (going to the mall): 3<br>bersepeda (cycling): 2<br>nonton tv (watch tv): 1<br>... | ID |
| 대한민국에서 주말에 아이와 함께하는 인기 있는 가족 활동은 무엇인가요? | 여행 (travel): 2<br>스포츠 (sports): 1<br>보드 게임 (board game): 1<br>... | KR |
| 북한에서 휴식일에 아이와 함께하는 많이 하는 가족 활동은 무엇인가요? | 사사끼 (card game): 1<br>장마당가기 (go to the market): 1<br>영화보기 (watching movie): 1<br>... | KP |
| Ποια είναι μια δημοφιλής οικογενειακή δραστηριότητα με ένα παιδί για τα σαββατοκύριακα στην Ελλάδα; | βόλτα (stroll): 1<br>κινηματογράφος (cinima): 1<br>παιδική χαρά (playground): 1 | GR |
| درایران یک فعالیت خانوادگی محبوب با فرزند برای انجام دادن در آخر هفته‌ها چیست؟ | پیک نیک در پارک (picnic in the park): 1<br>سفر (travel): 1<br>مهمانی (party): 1<br>... | IR |
| ما هي النشاطات العائلية الشائعة التي يمكن القيام بها مع الأطفال في عطلة نهاية الأسبوع في الجزائر؟ | التنزه (hiking): 5 | DZ |
| Azərbaycanda həftə sonları ailə ilə birlikdə uşaqla nə etmək populyardır? | parklara getmək (go to parks): 3<br>oyun meydançalarına getmək (go to playgrounds): 1<br>bağ evinə getmək (go to the country house): 1<br>... | AZ |
| Naon kagiatan kulawarga anu populer dipigawe babarengan jeung budak pikeun dilakukeun dina ahir minggu di Jawa Barat? | olahraga (sports): 1<br>lalajo tipi (watching tv): 1<br>ngojay (swimming):1<br>... | JB |
| অসমত সপ্তাহান্তত শিশুসহ পৰিয়ালে কি জনপ্ৰিয় কাম কৰে? | ফুৰিব যায় (go for a walk): 3<br>গাৰ্ডেনিং (gardening): 1<br>পিকনিকলৈ যায় (picnic): 1 | AS |
| Menene shahararren aikin gida da yara suka fi so suyi a karshen mako a Arewacin Najeriya? | shara (sweep): 3<br>wanki (washing): 1 | NG |
| በኢትዮጵያ በሳምንት መጨረሻ ቤተሰብ ከልጅ ጋር ለመስራት የታወቀ እንቅስቃሴ ምንድን ነው? | ሩጫ (running): 2<br>ልብስ ማጠብ (washing clothes): 1<br>ቤት ማጽዳት (house cleaning) | ET |

Figure 3: Example annotations for a cultural question related to the topic of *family* for each country/region in our dataset. The questions and annotations are provided in different languages, with translations of the annotated answers into English included in brackets. Annotations are sorted in descending order based on the frequency (i.e., vote count) of an answer provided by annotators, each separated by a line break. The vote count for each answer is displayed as numbers.



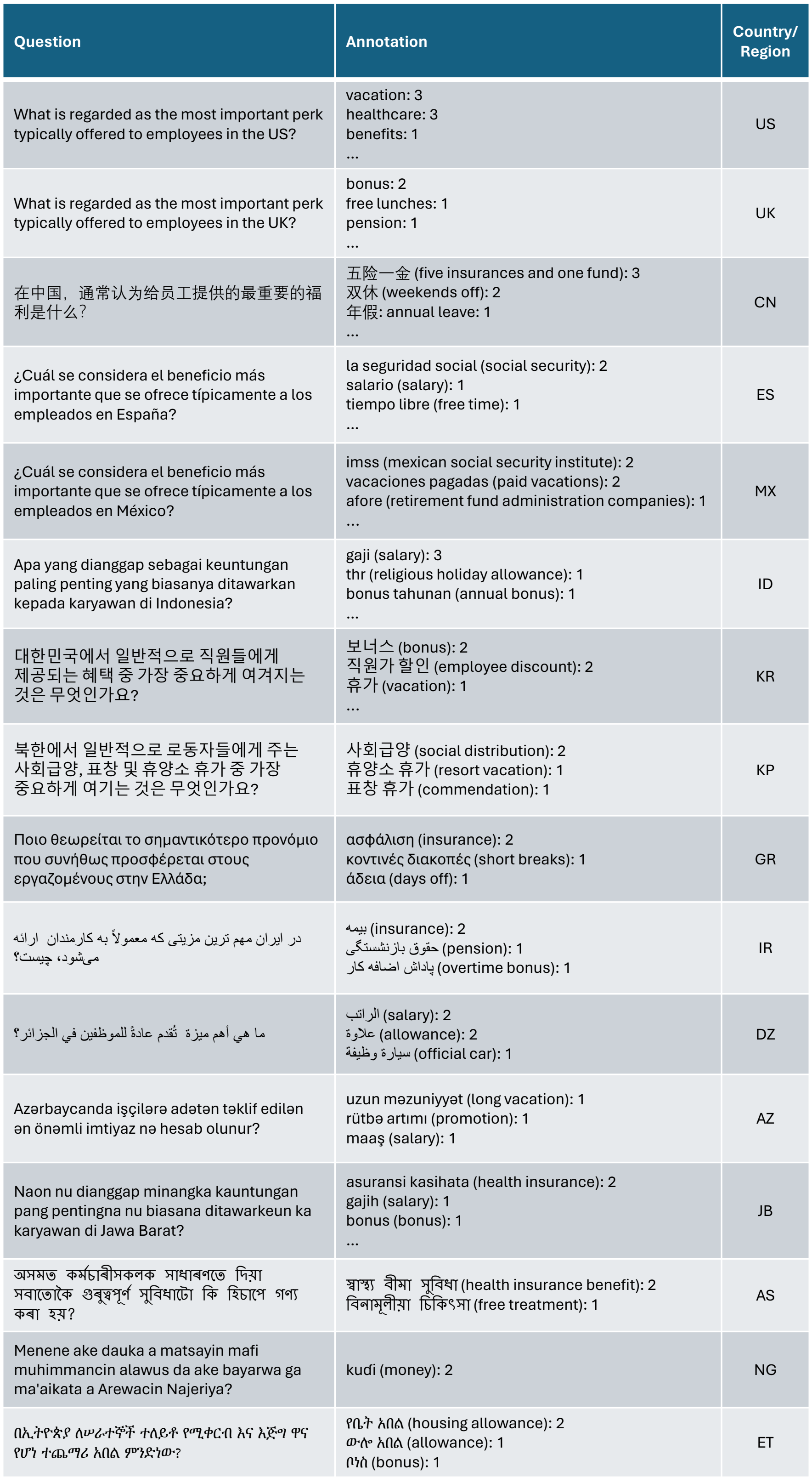
| Question | Annotation | Country/ Region |
|---|---|---|
| What language is taught in schools in the US besides English? | spanish: 5<br>french: 3<br>german: 2<br>... | US |
| What language is taught in schools in the UK besides English? | french: 5<br>spanish: 3<br>german: 2 | UK |
| 在中国的学校里除了英语之外还教授哪种语言？ | 中文 (chinese): 4 | CN |
| ¿Qué idioma se enseña en las escuelas de España además del inglés? | francés (french): 5<br>latin (latin): 2<br>aleman (german): 1<br>... | ES |
| ¿Qué idioma se enseña en las escuelas de México además del inglés? | francés (french): 4<br>español (spanish): 2<br>nahuatl (nahuatl): 1<br>... | MX |
| Bahasa apa yang diajarkan di sekolah-sekolah di Indonesia selain Bahasa Inggris? | bahasa indonesia (indonesian): 2<br>mandarin (mandarin): 2<br>bahasa daerah (regional language): 1<br>... | ID |
| 대한민국의 학교에서 학생들은 영어 외에 어떤 언어를 배우나요? | 일본어 (japanese): 4<br>중국어 (chinese): 3<br>불어 (french): 1 | KR |
| 북한의 학교에서 학생들은 영어 외에 어떤 외국어를 배우나요? | 중국어 (chinese): 4<br>러시아어 (russian language): 3<br>한문 (chinese characters): 1 | KP |
| Ποια γλώσσα διδάσκεται στα σχολεία στην Ελλάδα πέρα από τα Αγγλικά; | γερμανικά (german): 5<br>γαλλικά (french): 5<br>ελληνικά (greek): 1 | GR |
| درایران به جز انگلیسی، چه زبان‌هایی در مدارس تدریس داده می‌شود؟ | عربی (arabic): 4<br>انگلیسی (english): 1<br>فرانسه (france): 1<br>... | IR |
| أي لغة تُدرَّس في المدارس الجزائرية بالإضافة إلى اللغة الإنجليزية؟ | الفرنسية (french): 5 | DZ |
| Azərbaycanda məktəblərdə ingilis dilindən başqa hansı dillər tədris edilir? | rus dili (russian): 5<br>alman dili (german): 2<br>fransız dili (french): 1 | AZ |
| Basa naon nu diajarkeun di sakola-sakola di Jawa Barat salian ti Basa Inggris? | basa indonesia (indonesian language): 4<br>basa sunda (sundanese language): 2<br>jepang (japanese language):2<br>... | JB |
| অসমৰ বিদ্যালয়সমূহত ইংৰাজীৰ উপৰিও আন কোন ভাষা শিক্ষা দিয়া হয়? | हिन्दी (hindi): 5<br>সংস্কৃত (sanskrit): 2<br>অসমীয়া (assamese): 2<br>... | AS |
| Wane yare ake koyarwa a makarantun Arewacin Najeriya banda Turanci? | hausa (hausa): 4<br>larabci (arabic): 4 | NG |
| በኢትዮጵያ ትምህርት ቤቶች ከእንግሊዝኛ ቋንቋ በተጨማሪ ምን ይማራል? | አማርኛ (amharic): 4<br>ኦሮምኛ (oromic): 1 | ET |

Figure 4: Example annotations for a cultural question related to the topic of *educate* for each country/region in our dataset. The questions and annotations are provided in different languages, with translations of the annotated answers into English included in brackets. Annotations are sorted in descending order based on the frequency (i.e., vote count) of an answer provided by annotators, each separated by a line break. The vote count for each answer is displayed as numbers.



| Question | Annotation | Country/ Region |
|---|---|---|
| On which holiday do all family members tend to reunite in the US? | thanksgiving: 4<br>christmas: 2 | US |
| On which holiday do all family members tend to reunite in the UK? | christmas: 5 | UK |
| 在中国，哪个节日家里的所有成员会团聚? | 春节 (spring festival): 4<br>中秋节 (mid-autumn festival): 4<br>清明 (qingming): 1 | CN |
| ¿En qué festivo suelen reunirse todos los miembros de la familia en España? | navidad (christmas): 3<br>nochebuena (christmas eve): 2<br>nochevieja (new year's eve): 2<br>... | ES |
| ¿En qué festividad suelen reunirse todos los miembros de la familia en México? | navidad (christmas): 5<br>año nuevo (new year): 3<br>16 de septiembre (september 16th): 1<br>... | MX |
| Pada hari libur apa semua anggota keluarga biasanya berkumpul di Indonesia? | idul fitri (eid al-fitr): 4<br>natal (christmas):3<br>tahun baru (new year): 2<br>... | ID |
| 대한민국에서 모든 가족 구성원들이 함께 모이는 명절은 무엇이 있나요? | 추석 (chuseok): 5<br>설날 (lunar new year): 5 | KR |
| 북한에서 모든 가족 식구들이 함께 모이는 명절은 무엇이 있나요? | 추석 (chuseok): 3<br>설날 (lunar new year): 2<br>양력설 (gregorian new year): 1<br>... | KP |
| Σε ποια εορτή συνηθίζουν όλα τα μέλη της οικογένειας να επανασυνδέονται στην Ελλάδα; | πάσχα (easter): 4<br>χριστούγεννα (christmas): 3<br>γενέθλια (birthday): 1 | GR |
| در ایران در کدام تعطیلات همه اعضای خانواده معمولاً دور هم جمع می‌شوند؟ | نوروز (new year): 4<br>چهارشنبه سوری (chaharshanbe suri): 1<br>سیزده بدر (nature's day): 1<br>... | IR |
| في أي عيد يجتمع أفراد العائلة في الجزائر؟ | عيد الفطر (eid al-fitr): 5<br>عيد الاضحى (eid al-adha): 4<br>رأس السنة (new year): 1 | DZ |
| Azərbaycanda ailə üzvləri hansı bayramda bir araya gəlirlər? | novruz bayramı (novruz): 5<br>yeni il bayramı (new year): 1 | AZ |
| Dina liburan naon sadaya anggota kulawarga biasana ngariung deui di Jawa Barat? | idul fitri (eid al-fitr): 4<br>libur lebaran (eid holiday): 1<br>natal (christmas):1<br>... | JB |
| অসমত কোন উৎসৱত সকলো পৰিয়ালৰ সদস্যসকল একত্ৰিত হ'বলৈ প্ৰৱণ হয়? | बिहु (bihu): 5<br>पूजा (puja): 1<br>दुर्गा पूजा (durga puja): 2 | AS |
| A wane hutun ne dukkan 'yan uwa sukan hadu a Arewacin Najeriya? | hutun sallah (eid holiday): 4<br>hutun kistimeti (christmas): 3 | NG |
| በኢትዮጵያ በየትኛው በዓል ሁሉም ቤተሰቦች በአንድ ላይ ለመሆን ይሻሉ? | ፋሲካ (easter): 2<br>ረመዳን (ramadan): 1<br>ዘመን መለወጫ (new year) | ET |

Figure 5: Example annotations for a cultural question related to the topic of *holiday* for each country/region in our dataset. The questions and annotations are provided in different languages, with translations of the annotated answers into English included in brackets. Annotations are sorted in descending order based on the frequency (i.e., vote count) of an answer provided by annotators, each separated by a line break. The vote count for each answer is displayed as numbers.



| Question | Annotation | Country/ Region |
|---|---|---|
| What is regarded as the most important perk typically offered to employees in the US? | vacation: 3<br>healthcare: 3<br>benefits: 1<br>... | US |
| What is regarded as the most important perk typically offered to employees in the UK? | bonus: 2<br>free lunches: 1<br>pension: 1<br>... | UK |
| 在中国，通常认为给员工提供的最重要的福利是什么？ | 五险一金 (five insurances and one fund): 3<br>双休 (weekends off): 2<br>年假: annual leave: 1<br>... | CN |
| ¿Cuál se considera el beneficio más importante que se ofrece típicamente a los empleados en España? | la seguridad social (social security): 2<br>salario (salary): 1<br>tiempo libre (free time): 1<br>... | ES |
| ¿Cuál se considera el beneficio más importante que se ofrece típicamente a los empleados en México? | imss (mexican social security institute): 2<br>vacaciones pagadas (paid vacations): 2<br>afore (retirement fund administration companies): 1<br>... | MX |
| Apa yang dianggap sebagai keuntungan paling penting yang biasanya ditawarkan kepada karyawan di Indonesia? | gaji (salary): 3<br>thr (religious holiday allowance): 1<br>bonus tahunan (annual bonus): 1<br>... | ID |
| 대한민국에서 일반적으로 직원들에게 제공되는 혜택 중 가장 중요하게 여겨지는 것은 무엇인가요? | 보너스 (bonus): 2<br>직원가 할인 (employee discount): 2<br>휴가 (vacation): 1<br>... | KR |
| 북한에서 일반적으로 로동자들에게 주는 사회급양, 표창 및 휴양소 휴가 중 가장 중요하게 여기는 것은 무엇인가요? | 사회급양 (social distribution): 2<br>휴양소 휴가 (resort vacation): 1<br>표창 휴가 (commendation): 1 | KP |
| Ποιο θεωρείται το σημαντικότερο προνόμιο που συνήθως προσφέρεται στους εργαζομένους στην Ελλάδα; | ασφάλιση (insurance): 2<br>κοντινές διακοπές (short breaks): 1<br>άδεια (days off): 1 | GR |
| در ایران مهم ترین مزیتی که معمولاً به کارمندان  ارائه می‌شود، چیست؟ | بیمه (insurance): 2<br>حقوق بازنشستگی (pension): 1<br>پاداش اضافه کار (overtime bonus): 1 | IR |
| ما هي أهم ميزة  تُقدم عادةً للموظفين في الجزائر؟ | الراتب (salary): 2<br>علاوة (allowance): 2<br>سيارة وظيفة (official car): 1 | DZ |
| Azərbaycanda işçilərə adətən təklif edilən ən önəmli imtiyaz nə hesab olunur? | uzun məzuniyyət (long vacation): 1<br>rütbə artımı (promotion): 1<br>maaş (salary): 1 | AZ |
| Naon nu dianggap minangka kauntungan pang pentingna nu biasana ditawarkeun ka karyawan di Jawa Barat? | asuransi kasihata (health insurance): 2<br>gajih (salary): 1<br>bonus (bonus): 1<br>... | JB |
| অসমত কৰ্মচাৰীসকলক সাধাৰণতে দিয়া সবাতোকৈ গুৰুত্বপূৰ্ণ সুবিধাটো কি হিচাপে গণ্য কৰা হয়? | স্বাস্থ্য বীমা সুবিধা (health insurance benefit): 2<br>বিনামূলীয়া চিকিৎসা (free treatment): 1 | AS |
| Menene ake dauka a matsayin mafi muhimmancin alawus da ake bayarwa ga ma'aikata a Arewacin Najeriya? | kuɗi (money): 2 | NG |
| በኢትዮጵያ ለሠራተኞች ተለይቶ የሚቀርብ እና እጅግ ዋና የሆነ ተጨማሪ አበል ምንድነው? | የቤት አበል (housing allowance): 2<br>ውሎ አበል (allowance): 1<br>ቦነስ (bonus): 1 | ET |

Figure 6: Example annotations for a cultural question related to the topic of *work life* for each country/region in our dataset. The questions and annotations are provided in different languages, with translations of the annotated answers into English included in brackets. Annotations are sorted in descending order based on the frequency (i.e., vote count) of an answer provided by annotators, each separated by a line break. The vote count for each answer is displayed as numbers.

# 2 Construction Details of BLEnD

## 2.1 Resource Availability of Languages

As illustrated in the main text, we select languages with varying levels of resource availability and recruit annotators who are native speakers of each language. The detailed resource availability of the languages included in BLEnD is shown in Table 2.

Table 2: Resource availability of the 13 languages covered in BLEnD. The resource availability is defined by [5].

| Class | Languages |
|---|---|
| 1 - The Left-Behinds | Assamese, Azerbaijani, Sundanese |
| 2 - The Hopefuls | Amharic, Hausa |
| 3 - The Rising Stars | Greek, Indonesian |
| 4 - The Underdogs | Korean, Persian |
| 5 - The Winners | Arabic, Chinese (Mandarin), English, Spanish |

## 2.2 Ethical Considerations of Annotator Recruitment

This research project was performed under approval from KAIST IRB (KH2023-226). We obtained 'Informed Consent for Human Subjects' from the annotators. We embedded the consent document within the annotation website for the crowdworkers or received written consent from the directly recruited annotators. The annotations were gathered only from those who had read and consented to the form. We recruited annotators without any discrimination based on age, ethnicity, disability, or gender. Workers were compensated at a rate exceeding Prolific's ethical standards [1]. These same standards were applied to workers directly recruited for the annotation of low-resource languages.

Participants could voluntarily decide to join or withdraw from the study, and any data provided would not be used for research purposes if they withdraw. Additionally, the annotators were notified that if an unexpected situation arises during participation, appropriate actions will be taken according to the situation, and documents complying with the requirements of the KAIST IRB will be promptly prepared and reported.

## 2.3 Annotator Demographics

The statistics of all annotators participating in our dataset construction are shown in Table 3 and 4.

[1] https://www.prolific.com/resources/how-much-should-you-pay-research-participants

Table 3: Annotator demographics for each country or region who are recruited via Prolific.

| | US | GB | CN | ES | ID | GR | MX | IR |
|---|---|---|---|---|---|---|---|---|
| **No. of Annotators** | 87 | 119 | 59 | 91 | 40 | 86 | 86 | 50 |
| **Gender (%)** | | | | | | | | |
| Female | 42.53 | 46.22 | 55.93 | 49.45 | 50.00 | 45.35 | 48.84 | 56.00 |
| Male | 52.87 | 49.58 | 44.07 | 49.45 | 50.00 | 54.65 | 48.84 | 42.00 |
| Non-binary | 4.60 | 2.52 | - | 1.10 | - | - | 2.33 | 2.00 |
| Prefer not to say | - | 1.68 | - | - | - | - | - | - |
| **Age (%)** | | | | | | | | |
| -29 | 36.78 | 13.45 | 64.41 | 41.76 | 45.00 | 50.00 | 59.30 | 48.00 |
| 30-39 | 19.54 | 26.89 | 25.42 | 23.08 | 35.00 | 29.07 | 26.74 | 44.00 |
| 40-49 | 17.24 | 21.01 | 3.39 | 18.68 | 12.50 | 13.95 | 8.14 | 8.00 |
| 50-59 | 14.94 | 21.85 | 6.78 | 14.29 | 7.50 | 6.98 | 4.65 | - |
| 60+ | 11.49 | 16.81 | - | 2.20 | - | - | 1.16 | - |
| **Duration of Residence in Target Country (%)** | | | | | | | | |
| 100% | 55.17 | 75.63 | 1.69 | 75.82 | 5.00 | 86.05 | 75.58 | 8.00 |
| $\geq$ 90% | 9.20 | 7.56 | 28.81 | 10.99 | 25.00 | 1.16 | 16.28 | 34.00 |
| $\geq$ 80% | 13.79 | 5.04 | 23.73 | 5.49 | 20.00 | 6.98 | 2.33 | 22.00 |
| $\geq$ 70% | 6.90 | 3.36 | 15.25 | 5.49 | 17.50 | 5.81 | 4.65 | 20.00 |
| $\geq$ 60% | 9.20 | 5.04 | 25.42 | 2.20 | 12.50 | - | 1.16 | 10.00 |
| $\geq$ 50% | 5.75 | 2.52 | 5.08 | - | 20.00 | - | - | 6.00 |
| **Education Level (%)** | | | | | | | | |
| Below High School | - | 0.84 | - | 3.30 | - | - | - | 2.00 |
| High School | 11.49 | 12.61 | 6.78 | 12.09 | 20.00 | 13.95 | 15.12 | 4.00 |
| College | 22.99 | 21.85 | 3.39 | 16.48 | 2.50 | 11.63 | 4.65 | 10.00 |
| Bachelor | 47.13 | 48.74 | 35.59 | 40.66 | 30.00 | 40.70 | 66.28 | 32.00 |
| Master's Degree | 18.39 | 13.45 | 38.98 | 21.98 | 40.00 | 25.58 | 11.63 | 46.00 |
| Doctorate | - | 2.52 | 15.25 | 5.49 | 7.50 | 8.14 | 2.33 | 6.00 |

Table 4: Annotator demographics for each country or region who are recruited directly.

| | KR | DZ | AZ | KP | JB | AS | NG | ET |
|---|---|---|---|---|---|---|---|---|
| **No. of Annotators** | 5 | | | | | | | |
| **Gender (%)** | | | | | | | | |
| Female | 60.00 | 40.00 | 40.00 | 80.00 | 40.00 | 100.00 | 60.00 | - |
| Male | 40.00 | 60.00 | 60.00 | 20.00 | 60.00 | - | 40.00 | 100.00 |
| Non-binary | - | - | - | - | - | - | - | - |
| Prefer not to say | - | - | - | - | - | - | - | - |
| **Age (%)** | | | | | | | | |
| -29 | 60.00 | 20.00 | 100.00 | - | 100.00 | 60.00 | 60.00 | 60.00 |
| 30-39 | - | 60.00 | - | - | - | 40.00 | 40.00 | 40.00 |
| 40-49 | - | - | - | 40.00 | - | - | - | - |
| 50-59 | 40.00 | 20.00 | - | 60.00 | - | - | - | - |
| 60+ | - | - | - | - | - | - | - | - |
| **Duration of Residence in Target Country (%)** | | | | | | | | |
| 100% | 20.00 | 80.00 | - | - | 80.00 | 80.00 | 80.00 | 100.00 |
| $\geq$ 90% | - | - | - | - | - | - | - | - |
| $\geq$ 80% | 40.00 | - | 80.00 | 20.00 | - | - | 20.00 | - |
| $\geq$ 70% | 20.00 | 20.00 | 20.00 | - | 20.00 | - | - | - |
| $\geq$ 60% | 20.00 | - | - | - | - | - | - | - |
| $\geq$ 50% | - | - | - | 20.00 | - | 20.00 | - | - |
| $<$ 50% | - | - | - | 60.00 | - | - | - | - |
| **Education Level (%)** | | | | | | | | |
| Below High School | - | - | - | - | - | - | - | - |
| High School | 60.00 | - | 80.00 | - | 40.00 | - | 20.00 | - |
| College | - | - | - | 20.00 | - | - | - | - |
| Bachelor | 40.00 | 40.00 | 20.00 | 20.00 | 60.00 | 20.00 | 60.00 | 20.00 |
| Master's Degree | - | 40.00 | - | 60.00 | - | 80.00 | 20.00 | 80.00 |
| Doctorate | - | 20.00 | - | - | - | - | - | - |

Table 5: Average of maximum votes among all answers for each question in different categories across countries. A value of ‘3.00’ indicates that, on average, three annotators provided the same answer for each question.

| **Category** | **US** | **GB** | **ES** | **MX** | **ID** | **CN** | **KR** | **DZ** | **GR** | **IR** | **KP** | **AZ** | **JB** | **AS** | **NG** | **ET** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Food | 3.12 | 3.14 | 2.99 | 2.67 | 2.93 | 3.27 | 3.28 | 3.29 | 2.91 | 2.99 | 2.61 | 3.19 | 3.01 | 3.14 | 2.72 | 3.04 |
| Sport | 3.35 | 3.47 | 3.57 | 3.07 | 3.59 | 3.53 | 3.57 | 3.09 | 3.30 | 3.59 | 2.89 | 3.24 | 3.47 | 2.97 | 2.98 | 3.18 |
| Family | 3.17 | 3.40 | 3.17 | 3.08 | 3.16 | 3.16 | 3.40 | 2.94 | 3.19 | 3.17 | 2.81 | 3.25 | 2.94 | 3.19 | 2.65 | 2.78 |
| Education | 3.24 | 3.26 | 3.30 | 3.25 | 3.21 | 3.19 | 3.63 | 3.18 | 3.29 | 3.20 | 3.27 | 3.42 | 3.45 | 3.10 | 2.94 | 3.23 |
| Holidays | 3.09 | 3.33 | 3.18 | 3.04 | 3.14 | 3.28 | 3.60 | 3.04 | 2.98 | 3.20 | 3.07 | 3.27 | 3.10 | 2.92 | 2.60 | 3.12 |
| Work-life | 3.10 | 3.19 | 3.09 | 3.15 | 3.22 | 3.00 | 3.57 | 3.31 | 2.87 | 3.09 | 3.01 | 3.59 | 3.10 | 3.25 | 2.75 | 3.12 |
| **Overall** | **3.18** | **3.29** | **3.22** | **3.02** | **3.20** | **3.25** | **3.50** | **3.15** | **3.08** | **3.21** | **2.93** | **3.31** | **3.18** | **3.08** | **2.78** | **3.09** |

### 2.4 Question Construction Guidelines

Below are the annotation guidelines for creating the question templates in BLEnD.

The goal of this task is to write question-and-answer pairs that ask about your country’s culture. In each spreadsheet, you need to write down the questions and the corresponding answers to each question. Write them down in your native language, and add their translation into English too in the spreadsheet provided.

Please find below a few guidelines to take into account when writing the questions:

- **Questions and answers should be a culture specific question related to your culture** (can be a common sense question). For example, a question related to the sport topic could be “What is the most popular sport in your country?”. You should refrain from writing factual questions as much as possible.
- **Do not generate yes or no questions or answers that only have two options** (e.g. male or female). You could convert a yes or no question to a question starting with question words. Instead of asking “‘Do people in your country tend to get off work at 5:30 pm?", you may ask “What time do people in your country tend to get off work?".
- **Please write questions distinct from each other as much as possible** under each topic.
- **The answer should be short and concrete**. It is better to use precise concepts, entities, time, etc. to answer each question.
- **Please avoid asking questions about a very stereotypical topic**. For instance, avoid questions like “Who bears more responsibility for taking care of children at home in your country?"

### 2.5 Answer Annotation Guidelines

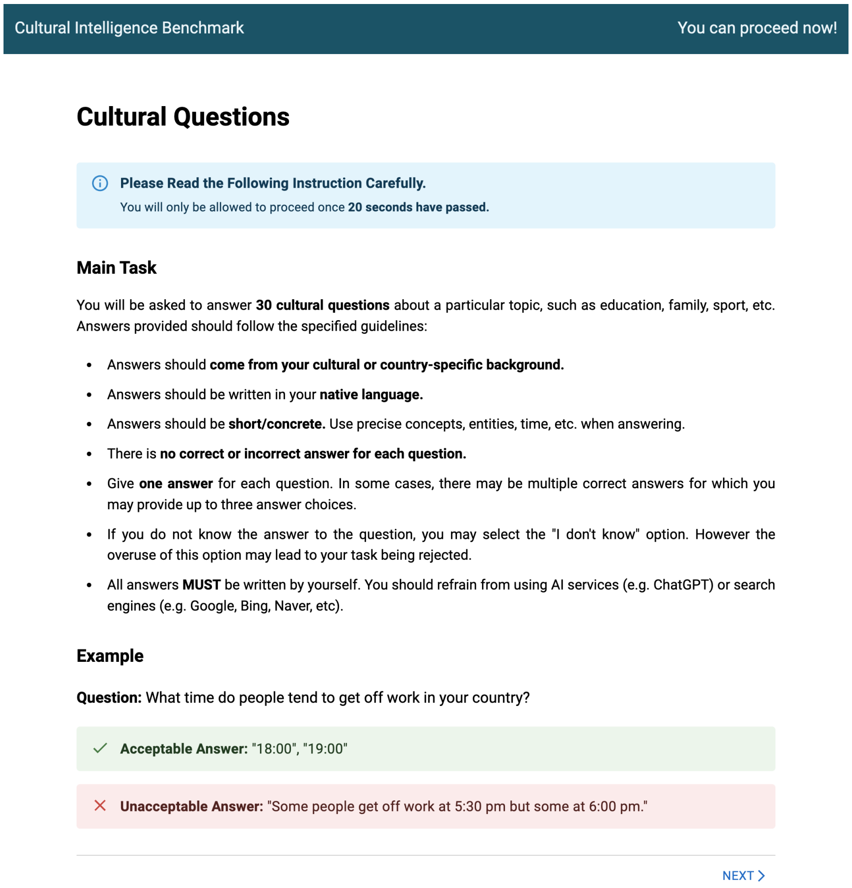
Figure 7 shows the annotation guidelines given to the annotators for all countries/regions. We provided guidelines, all in their local languages.

### 2.6 Answer Annotation Interface

Figure 8 shows the annotation interface shown to the crowdworkers annotators in Prolific. We used an Excel sheet for annotators recruited by direct recruitment for the annotations (i.e., for low-resource languages).

### 2.7 Annotation Analysis

Table 5 shows the level of agreement between the annotators, calculated by averaging the maximum votes among answers for each question in different categories across countries.



Cultural Intelligence Benchmark

You can proceed now!

**Cultural Questions**

**Please Read the Following Instruction Carefully.**

You will only be allowed to proceed once **20 seconds have passed.**

**Main Task**

You will be asked to answer **30 cultural questions** about a particular topic, such as education, family, sport, etc. Answers provided should follow the specified guidelines:

- Answers should **come from your cultural or country-specific background.**
- Answers should be written in your **native language.**
- Answers should be **short/concrete.** Use precise concepts, entities, time, etc. when answering.
- There is **no correct or incorrect answer for each question.**
- Give **one answer** for each question. In some cases, there may be multiple correct answers for which you may provide up to three answer choices.
- If you do not know the answer to the question, you may select the "I don't know" option. However the overuse of this option may lead to your task being rejected.
- All answers **MUST** be written by yourself. You should refrain from using AI services (e.g. ChatGPT) or search engines (e.g. Google, Bing, Naver, etc).

**Example**

**Question:** What time do people tend to get off work in your country?

✓ **Acceptable Answer:** "18:00", "19:00"

✗ **Unacceptable Answer:** "Some people get off work at 5:30 pm but some at 6:00 pm."

Figure 7: Answer annotation guidelines shown to the annotators.

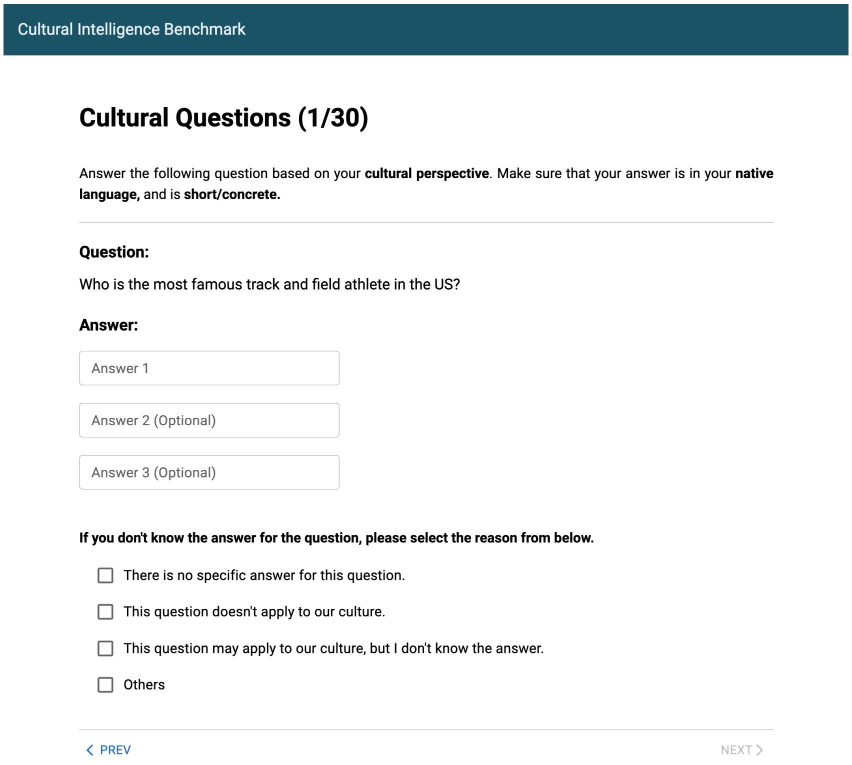
# 3 Experimental Settings for LLM Evaluation

## 3.1 Models

We use GPT-4 (`gpt-4-1106-preview`), GPT-3.5 (`gpt-3.5-turbo-1106`)[2], Claude-3-Opus (`claude-3-opus-20240229`), Claude-3-Sonnet (`claude-3-sonnet-20240229`), Claude-3-Haiku (`claude-3-haiku-20240307`)[3], PaLM2 (`text-bison-002`)[4], Gemini-1.0-Pro[5], C4AI Command R+[6], C4AI Command R[7], Qwen-1.5-72B/32B/14B-Chat [2], SeaLLM-7B-v2.5 [7], Hyper CLOVA X [9], Aya-23 [1] and Aya-101 [11] for evaluation of cultural sensitivity of LLMs.

8 Quadro RTX 8000 48GB were used with CUDA version 11.8 for all experiments. For all models, we use greedy decoding (temperature=0, top_p=1.0). We use the PyTorch library[8] for all experiments.

[2] `https://platform.openai.com/docs/models`
[3] `https://www.anthropic.com/api`
[4] `https://cloud.google.com/vertex-ai/generative-ai/docs/model-reference/text`
[5] `https://ai.google.dev/gemini-api/docs/models/gemini?hl=ko`
[6] `https://docs.cohere.com/reference/about`
[7] `https://huggingface.co/CohereForAI/c4ai-command-r-v01`
[8] `https://pytorch.org/`



Cultural Intelligence Benchmark

## Cultural Questions (1/30)

Answer the following question based on your **cultural perspective**. Make sure that your answer is in your **native language,** and is **short/concrete.**

**Question:**

Who is the most famous track and field athlete in the US?

**Answer:**

Answer 1

Answer 2 (Optional)

Answer 3 (Optional)

**If you don't know the answer for the question, please select the reason from below.**

□ There is no specific answer for this question.

□ This question doesn't apply to our culture.

□ This question may apply to our culture, but I don't know the answer.

□ Others

Figure 8: Annotation interface given to the annotators.

## 3.2 Short Answer Question

### 3.2.1 Prompt for Short Answer Evaluation

We use two types of prompts to evaluate the LLMs in short answer settings: 1) directly ask LLMs to provide the answer and 2) add persona to the LLMs to make them act as a person from the target country or region. The details of each prompt are as follows:

PROMPT #1:
Read the following question and provide a single answer without any explanations.

Question: {QUESTION}
Answer:

PROMPT #2:
You are a person from {COUNTRY/REGION} who is trying to explain your country's culture to a foreigner. Answer the following question, providing a single answer without any explanations.

{QUESTION}

### 3.2.2 Details of Short Answer Evaluation

Let $Q$ denote the question set, $A_q$ the annotated answer set for each question $q \in Q$, with each answer $a \in A_q$, for a question $q$ in the country or region $c$ in the human annotation. For any LLM prediction $y$, we define $s_{q,c}(y)$ as

$$s_{q,c}(y) = \begin{cases} 1, & \text{if } \exists a \in A_q \text{ such that } a \subseteq y \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

so that $s_{q,c}(y)$ is 1 if the prediction $y$ includes any of the answers from the human annotations, denoted as $a \subseteq y$, and 0 otherwise. For a model $m$ that outputs $f_m(q, c)$ when given $q$ and $c$, the score $S(c)$ for each country or region $c$ is calculated as

$$S(c) = \frac{1}{|Q|} \sum_{q \in Q} s_{q,c}(f_m(q,c)) \times 100. \tag{2}$$

To evaluate LLM responses, we lemmatize/stem/tokenize the annotations and LLM responses for each question to consider the language variations. We use one of the three techniques that are available for each language.

We use the lemmatizer from the English model from SpaCy (`en_core_web_sm`) for English. For Spanish and Amharic, we use lemmatizers from SparkNLP [9]. For Indonesian, we use the lemmatizer from Kumparan NLP Library [10]. For Chinese, we use jieba [11], a Chinese word segmentation module. For Korean, we use the Okt lemmatizer from the konlpy package [12]. For Arabic, we use Qalsadi Arabic Lemmatizer [10]. For Greek, we use the CLTK Greek lemmatizer [4]. For Persian, we use Hazm, a Persian NLP Toolkit [13]. For Azerbaijani, we use the Azerbaijani Language Stemmer [14]. We use SUSTEM, a Sundanese Stemmer [8] for Sundanese. We use the Assamese tokenizer from Indic NLP Library [6] for Assamese. For Hausa, we use the Hausa Stemmer [3].

## 3.3 Multiple Choice Question

### 3.3.1 Multiple Choice Question Construction

To create plausible incorrect answer options for questions about the target country/region, we first consider all answer annotations from all other countries with at least two votes. Then, we sort these answer candidates by their vote count from each country/region. Next, we check each candidate to see if it is similar to any annotations collected from the target country/region. If it is, we block that candidate from being added as a wrong answer choice, as well as the same answer from the other countries/regions. We use GPT-4 to determine if two words are similar in meaning, such as ‘fruit’ and ‘apple’, as the two can be considered the same when answering the question. The prompt can be seen in Appendix 3.3.2.

As this process would lead to differing possible wrong answer options for each target country per question, we pick the answer options with the minimum number of possible wrong answer options among all countries. If there are $n$ possible answer choices, we include all combinations of $\binom{n}{3}$ if $n \geq 3$, or include all $n$ answer choices plus $3 - n$ dummy options otherwise. We use GPT-4 (see Appendix 3.3.2 for the prompt details) to produce dummy answer options to make the number of options comprised of one correct answer and three wrong answer options four. If there are multiple correct answers, we generate multiple versions of the question, each with a different correct answer. The choices are provided in alphabetical order when asked to LLMs in a multiple-choice format.

### 3.3.2 Prompt for Multiple Choice Question Construction

**Similar Term Detection.** Since we asked the human annotators to provide answers in a short answer format, there may be cases where different textual answers refer to the same meaning. To avoid duplicate options in multiple-choice format, we utilized GPT-4 to determine whether the answers have the same meaning using the following prompt:

> Determine if a ‘target’ word is the same in meaning(e.g., football & soccer or soccer & football)

[9] Spanish lemmatizer (`https://sparknlp.org/2020/02/16/lemma_es.html`), Amharic lemmatizer (`https://sparknlp.org/2021/01/20/lemma_am.html`)

[10] `https://github.com/kumparan/nlp-id/tree/v0.1.9.9`

[11] `https://github.com/fxsjy/jieba?tab=readme-ov-file`

[12] `https://konlpy.org/en/latest/api/konlpy.tag/`

[13] `https://github.com/roshan-research/hazm`

[14] `https://github.com/aznlp-disc/stemmer`

> to at least one of the ‘answer’ words, or one is a subset to another(e.g., fruit & apple or apple & fruit). If so, the ‘result’ for ‘target’ word is ‘O’. However, if the two simply falls into the same level of hierarchy, the ‘result’ is ‘X’ (banana & apple, rose & carnation).
>
> Note that the ‘answer’ list is from ‘answer_country,’ and the ‘target’ word is from ‘target_country,’ as written by a person.
>
> Write down your reasoning first. Do not write any other JSON formatted object in your answer except for the result JSON object, formatted as {“result”:“O”} or {“result”:“X”}.

**Dummy Options Generation.** In cases where a question has fewer than four options during the option generation process, we ask GPT-4 to produce dummy options using the following prompt:

> Provide {$3 - n$} dummy option(s) that makes sense to be the answer(s) of the given “question”, and has to exist in real-life (non-fiction), but is totally different from the given “answers” without any explanation. Make sure that the options are different from each other, and cannot be an answer from any country. Provide as JSON format: {“dummy_options”:[]}

### 3.3.3 Prompt for Multiple Choice Evaluation

We use the following prompt to evaluate the LLMs’ performance in multiple-choice format:

> {QUESTION} Without any explanation, choose only one from the given alphabet choices(e.g., A, B, C). Provide as JSON format: {“answer_choice”:“”}
>
> A. {CHOICE 1}
> B. {CHOICE 2}
> C. {CHOICE 3}
> D. {CHOICE 4}
>
> Answer:

# 4 Detailed LLM Performance Analysis

## 4.1 LLM Evaluation Results

Table 6 and Table 7 show the performance of all LLMs experimented on the short answer questions for all countries/regions on the local language and English, respectively. Table 8 shows the performance of all LLMs on the multiple-choice questions for all countries/regions.

Table 6: Performance of all LLMs on short answer questions for each country/region in local language.

| | US | GB | ES | MX | ID | CN | KR | DZ | GR | IR | KP | AZ | JB | AS | NG | ET |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **en** | **en** | **es** | **es** | **id** | **zh** | **ko** | **ar** | **el** | **fa** | **ko** | **az** | **su** | **as** | **ha** | **am** |
| **GPT-4** | 83.19 | 82.75 | 79.00 | 77.45 | 77.50 | 77.32 | 80.95 | 67.62 | 70.43 | 73.03 | 49.32 | 62.05 | 55.79 | 49.06 | 45.93 | 25.85 |
| **Claude-3-Opus** | 83.84 | 78.79 | 78.78 | 75.57 | 78.02 | 76.90 | 78.95 | 65.68 | 69.24 | 77.85 | 55.41 | 69.62 | 56.55 | 52.41 | 46.37 | 35.38 |
| **Claude-3-Sonnet** | 81.34 | 81.65 | 72.60 | 72.44 | 75.73 | 66.77 | 66.32 | 61.33 | 63.48 | 67.32 | 45.05 | 59.28 | 45.09 | 38.89 | 27.14 | 26.59 |
| **Gemini-1.0-Pro** | 80.48 | 78.57 | 74.95 | 72.55 | 72.71 | 70.36 | 65.26 | 62.01 | 64.78 | 38.82 | 43.47 | 44.24 | 44.87 | 27.99 | 35.82 | 18.86 |
| **Command R+** | 80.48 | 78.35 | 73.67 | 70.77 | 72.19 | 64.87 | 75.05 | 62.13 | 59.89 | 67.11 | 49.55 | 41.15 | 31.22 | 25.89 | 16.26 | 5.51 |
| **Claude-3-Haiku** | 80.48 | 77.91 | 71.22 | 72.03 | 70.73 | 62.55 | 66.63 | 57.32 | 63.37 | 59.98 | 41.67 | 54.58 | 43.01 | 34.17 | 24.07 | 21.82 |
| **GPT-3.5** | 81.45 | 81.87 | 74.63 | 71.92 | 73.12 | 68.78 | 65.16 | 58.70 | 57.17 | 55.48 | 40.09 | 44.35 | 32.31 | 6.92 | 19.34 | 3.71 |
| **PaLM2** | 80.37 | 77.36 | 72.92 | 71.82 | 75.31 | 70.57 | 63.89 | 63.62 | 67.39 | 27.63 | 41.67 | 29.42 | 44.76 | 18.03 | 19.78 | 9.00 |
| **Qwen1.5-72B** | 83.95 | 79.34 | 70.04 | 70.15 | 65.31 | 78.27 | 60.53 | 54.81 | 32.93 | 39.25 | 38.96 | 36.89 | 32.42 | 18.45 | 9.67 | 8.90 |
| **SeaLLM** | 80.80 | 80.11 | 67.80 | 69.52 | 63.75 | 64.77 | 52.95 | 49.54 | 41.96 | 48.79 | 39.64 | 39.02 | 28.38 | 15.72 | 22.64 | 5.40 |
| **HyperCLOVA X** | 81.45 | 79.34 | 69.08 | 72.13 | 65.52 | 58.44 | 79.05 | 29.98 | 35.54 | 30.48 | 52.03 | 27.72 | 40.39 | 5.77 | 10.22 | 1.48 |
| **Qwen1.5-32B** | 82.43 | 79.67 | 59.70 | 60.65 | 58.44 | 79.11 | 52.74 | 41.53 | 35.33 | 44.08 | 33.22 | 35.71 | 26.31 | 22.22 | 11.21 | 4.87 |
| **Command R** | 77.87 | 77.58 | 68.55 | 66.81 | 63.02 | 60.76 | 60.84 | 57.78 | 54.78 | 59.98 | 40.54 | 9.70 | 29.04 | 13.52 | 11.65 | 3.18 |
| **Aya-23** | 77.33 | 72.09 | 69.62 | 66.81 | 69.58 | 62.03 | 66.84 | 55.38 | 58.15 | 59.32 | 43.24 | 27.40 | 25.44 | 8.49 | 5.16 | 3.07 |
| **Qwen1.5-14B** | 78.74 | 76.59 | 56.82 | 63.26 | 54.17 | 76.79 | 52.21 | 39.82 | 20.54 | 28.51 | 33.78 | 34.01 | 22.60 | 17.82 | 9.12 | 3.28 |
| **Aya-101** | 53.36 | 48.02 | 45.84 | 46.03 | 41.88 | 32.17 | 32.84 | 33.64 | 27.72 | 34.87 | 23.09 | 35.82 | 27.51 | 4.40 | 24.51 | 17.80 |

Table 7: Performance of all LLMs on short answer questions for each country/region in English.

| | **CN** | **ID** | **ES** | **GR** | **MX** | **KR** | **AZ** | **IR** | **DZ** | **AS** | **JB** | **KP** | **ET** | **NG** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **GPT-4** | 70.89 | 70.00 | 67.91 | 68.70 | 63.15 | 69.68 | 64.61 | 65.46 | 64.76 | 54.09 | 55.68 | 46.62 | 45.97 | 37.69 |
| **Claude-3-Opus** | 66.98 | 62.81 | 61.30 | 61.09 | 58.35 | 64.42 | 60.66 | 61.29 | 57.78 | 48.74 | 50.76 | 42.00 | 40.78 | 34.95 |
| **Claude-3-Sonnet** | 66.88 | 66.67 | 60.45 | 60.98 | 57.93 | 63.47 | 61.30 | 57.35 | 54.92 | 50.94 | 50.11 | 41.10 | 42.06 | 35.71 |
| **Gemini-1.0-Pro** | 66.46 | 59.27 | 59.70 | 60.54 | 56.47 | 59.68 | 57.46 | 55.92 | 53.78 | 44.55 | 49.89 | 42.68 | 40.15 | 32.42 |
| **Command R+** | 64.98 | 59.58 | 59.06 | 58.59 | 61.06 | 59.89 | 56.50 | 54.28 | 56.86 | 48.43 | 46.40 | 43.58 | 40.78 | 33.52 |
| **Claude-3-Haiku** | 60.44 | 59.38 | 53.62 | 56.52 | 55.74 | 59.89 | 56.29 | 53.18 | 52.29 | 45.70 | 46.18 | 37.84 | 35.49 | 34.40 |
| **GPT-3.5** | 64.66 | 63.23 | 62.26 | 61.85 | 61.48 | 60.00 | 59.59 | 56.36 | 57.67 | 48.43 | 49.56 | 44.48 | 40.04 | 38.46 |
| **PaLM2** | 66.14 | 62.19 | 60.45 | 60.98 | 58.14 | 60.00 | 57.68 | 55.92 | 56.29 | 47.38 | 48.47 | 43.36 | 38.03 | 33.08 |
| **Qwen1.5-72B** | 66.88 | 63.54 | 63.33 | 61.96 | 61.48 | 56.53 | 59.06 | 56.91 | 57.55 | 49.79 | 47.60 | 41.89 | 43.75 | 38.90 |
| **SeaLLM** | 65.61 | 62.81 | 62.58 | 59.46 | 60.44 | 56.95 | 58.42 | 60.20 | 52.97 | 51.78 | 48.69 | 41.89 | 42.90 | 43.08 |
| **HyperCLOVA X** | 62.76 | 63.65 | 67.06 | 60.33 | 63.05 | 62.74 | 56.61 | 56.91 | 55.15 | 51.68 | 50.76 | 44.03 | 45.34 | 40.22 |
| **Qwen1.5-32B** | 69.30 | 58.75 | 61.73 | 58.59 | 60.96 | 56.74 | 54.69 | 54.06 | 49.89 | 47.69 | 44.65 | 39.41 | 41.31 | 39.01 |
| **Command R** | 61.50 | 57.40 | 58.64 | 56.20 | 57.41 | 56.11 | 51.39 | 50.99 | 55.26 | 45.70 | 42.03 | 41.67 | 38.67 | 35.05 |
| **Aya-23** | 56.65 | 53.33 | 54.90 | 54.02 | 51.98 | 49.05 | 48.72 | 50.77 | 47.83 | 44.34 | 42.90 | 36.26 | 34.11 | 29.78 |
| **Qwen1.5-14B** | 64.66 | 55.73 | 55.12 | 52.83 | 60.44 | 54.53 | 51.92 | 52.96 | 48.51 | 45.39 | 40.94 | 33.00 | 39.72 | 39.89 |
| **Aya-101** | 34.28 | 38.65 | 35.71 | 38.04 | 38.52 | 30.74 | 31.88 | 28.95 | 30.89 | 34.70 | 28.49 | 24.32 | 26.38 | 23.41 |

Table 8: Performance of all LLMs on multiple-choice questions for each country/region in English.

| | **GB** | **US** | **CN** | **ES** | **MX** | **DZ** | **GR** | **KR** | **JB** | **IR** | **ID** | **AZ** | **KP** | **NG** | **AS** | **ET** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **GPT-4** | 94.17 | 93.34 | 93.70 | 92.04 | 87.98 | 89.28 | 86.73 | 88.10 | 87.90 | 86.49 | 87.81 | 86.58 | 78.59 | 76.40 | 71.79 | 66.52 |
| **Claude-3-Opus** | 95.74 | 93.18 | 93.05 | 91.52 | 89.19 | 85.98 | 84.75 | 86.83 | 85.41 | 87.39 | 81.36 | 85.81 | 74.93 | 77.32 | 74.99 | 64.78 |
| **Qwen1.5-72B** | 91.80 | 92.29 | 88.54 | 85.43 | 81.14 | 79.42 | 80.93 | 76.94 | 78.62 | 78.14 | 78.94 | 75.67 | 75.95 | 67.82 | 64.42 | 61.63 |
| **Qwen1.5-32B** | 91.94 | 89.79 | 89.98 | 84.45 | 79.26 | 76.09 | 80.40 | 72.31 | 74.75 | 76.54 | 74.33 | 72.95 | 72.71 | 71.72 | 64.04 | 61.00 |
| **Gemini-1.0-Pro** | 87.87 | 89.18 | 86.97 | 82.53 | 80.68 | 79.09 | 78.92 | 80.58 | 80.32 | 75.13 | 73.63 | 77.22 | 67.94 | 65.04 | 66.33 | 56.99 |
| **Claude-3-Sonnet** | 83.98 | 86.18 | 86.54 | 81.12 | 82.75 | 78.02 | 77.30 | 81.79 | 77.53 | 77.69 | 76.31 | 73.54 | 71.33 | 66.26 | 68.40 | 55.20 |
| **Command R+** | 85.16 | 83.03 | 79.46 | 80.18 | 77.23 | 76.00 | 78.39 | 73.06 | 78.10 | 77.12 | 79.15 | 72.56 | 64.92 | 70.65 | 61.94 | 64.69 |
| **PaLM2** | 89.38 | 86.75 | 83.18 | 79.10 | 77.24 | 79.68 | 76.96 | 73.02 | 78.37 | 72.94 | 73.69 | 73.72 | 64.10 | 66.46 | 66.75 | 57.53 |
| **GPT-3.5** | 86.87 | 88.83 | 80.30 | 82.37 | 78.74 | 76.64 | 75.54 | 71.10 | 74.93 | 72.78 | 72.03 | 74.13 | 63.34 | 71.73 | 61.54 | 64.22 |
| **Claude-3-Haiku** | 87.41 | 81.75 | 79.79 | 79.34 | 73.22 | 78.47 | 76.24 | 75.21 | 74.39 | 72.56 | 71.26 | 69.91 | 67.22 | 68.96 | 63.93 | 58.28 |
| **SeaLLM** | 82.66 | 83.17 | 80.08 | 76.41 | 71.78 | 72.68 | 74.29 | 74.71 | 65.14 | 70.84 | 72.24 | 71.15 | 60.93 | 67.41 | 58.99 | 58.83 |
| **Aya-23** | 82.45 | 79.83 | 79.47 | 76.24 | 72.17 | 72.36 | 70.90 | 71.49 | 71.82 | 70.56 | 72.52 | 67.51 | 62.98 | 63.59 | 55.42 | 54.32 |
| **Qwen1.5-14B** | 82.96 | 81.36 | 79.78 | 75.47 | 75.24 | 73.96 | 68.89 | 71.10 | 67.43 | 69.96 | 66.33 | 67.31 | 66.55 | 65.05 | 56.14 | 53.79 |
| **Command R** | 79.75 | 73.44 | 76.57 | 73.80 | 70.18 | 72.66 | 69.99 | 70.05 | 68.96 | 70.26 | 70.21 | 62.32 | 61.65 | 60.76 | 55.66 | 55.24 |
| **HyperCLOVA X** | 79.80 | 79.78 | 74.85 | 71.34 | 69.14 | 67.91 | 68.67 | 71.15 | 68.73 | 62.84 | 69.64 | 68.78 | 62.78 | 57.60 | 60.82 | 46.04 |
| **Aya-101** | 68.75 | 64.86 | 61.09 | 61.68 | 60.16 | 57.96 | 56.60 | 56.46 | 53.59 | 55.17 | 55.19 | 58.19 | 54.92 | 43.88 | 45.08 | 45.49 |

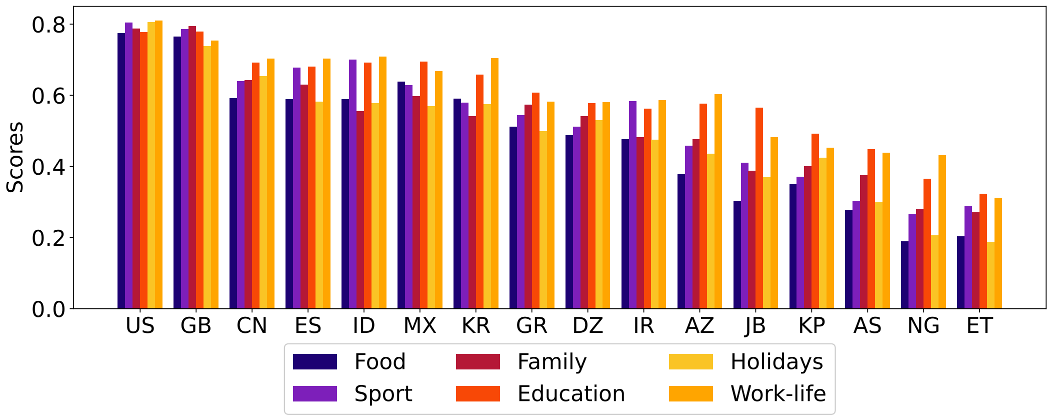


Figure 9: Average performance on all LLMs across all countries on each question category.

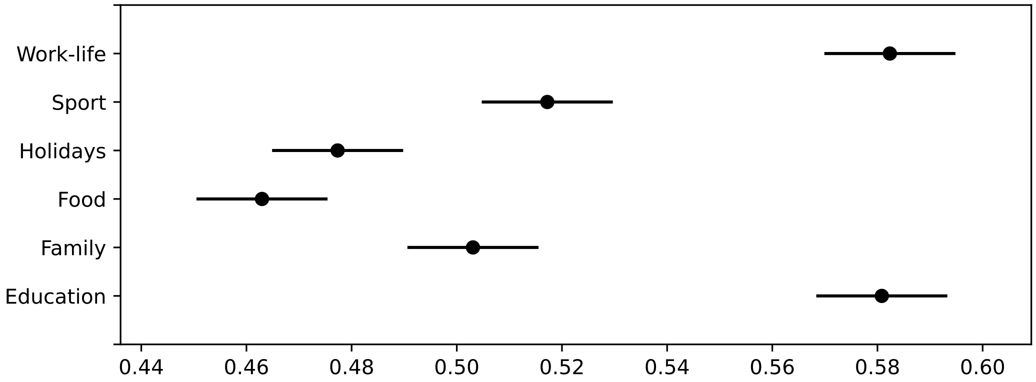


Figure 10: Tukey-HSD test on the LLM performances on each question category with 95% confidence interval.

## 4.2 LLM Performance by Question Category

Figure 9 illustrates the average performance of all LLMs for each category per country. This indicates that LLMs generally perform better in high-resource languages and countries. However, there are discrepancies in performance across different categories. LLMs do better on work-life or education-related questions but struggle with food and holidays/celebrations/leisure-related questions. This could be because the latter topics are more subjective. Figure 10 displays the results of the Tukey-HSD test on LLM performances for each topic, confirming that the performance difference between these two groups is statistically significant.

## 4.3 Human Evaluation

### 4.3.1 Human Evaluation Schema

The human evaluation is conducted on the following categories, which were decided based on the pilot annotations by the authors.

**Applicability.** We ask annotators to evaluate whether the LLM's response is applicable to the general population of their country/region. Since we take annotations from only 5 people per question, a correct answer from the annotator may not necessarily represent the whole culture and vice versa.

The applicability of the response is evaluated on three categories: 1) Applicable, 2) Conditionally Applicable, and 3) Incorrect. A response is annotated as applicable if all the answers provided by

Table 9: Summary of the human evaluation results across all countries. Scores are calculated by giving a weight of 1 for applicable, 0.5 for conditionally applicable, and 0 for incorrect responses. The values are presented as percentages, calculated by the number of responses that satisfy the criteria divided by the total number of responses. The country with the highest percentage is marked in **bold**, and the second highest is underlined.

| Country/Region | Score | Unnatural Language | Stereotypical | Partially Correct | Refusal | Nonsensical | Different Country's View |
|---|---|---|---|---|---|---|---|
| US | 66.67 | 3.33 | 0.83 | 0.00 | 4.17 | 5.83 | 2.50 |
| GB | **82.50** | 0.83 | 0.83 | 0.00 | 0.00 | 6.67 | 5.00 |
| ES | 39.17 | 0.00 | 1.67 | 5.00 | 0.00 | 10.00 | 11.67 |
| CN | 63.33 | 0.00 | 3.33 | 7.50 | 7.50 | 3.33 | 1.67 |
| ID | 60.00 | 0.83 | 13.33 | 2.50 | 1.67 | 18.33 | 4.17 |
| MX | 68.75 | 0.83 | 5.83 | 4.17 | 0.83 | 3.33 | 6.67 |
| KR | 50.42 | 0.83 | 7.50 | 3.33 | 8.33 | 5.00 | 8.33 |
| DZ | 47.50 | 0.00 | 14.17 | 8.33 | 2.50 | 7.50 | 6.67 |
| GR | 56.25 | 0.83 | 7.50 | 0.83 | 8.33 | 15.00 | 8.33 |
| IR | 56.67 | 0.00 | 13.33 | 10.83 | 2.50 | 10.00 | 0.00 |
| KP | 38.33 | **18.33** | 12.50 | 1.67 | 16.67 | 6.67 | 12.50 |
| AZ | 42.50 | 10.00 | 13.33 | 0.83 | **17.50** | 10.83 | **13.33** |
| JB | 44.58 | 6.67 | 21.67 | 5.00 | 3.33 | **38.33** | 1.67 |
| AS | 45.83 | 5.00 | 19.17 | 10.00 | 6.67 | 20.83 | 1.67 |
| NG | 36.25 | 7.50 | 2.50 | **22.50** | 0.83 | 18.33 | 7.50 |
| ET | 27.92 | 1.67 | **48.33** | 15.83 | 8.33 | 24.17 | 4.17 |

the model are valid for the general population of the country/region. When the response contains an answer that makes sense in some contexts but not necessarily to most people from the country/region, it is annotated as conditionally applicable. Finally, if at least one answer is completely inapplicable to the country/region, the response is annotated as incorrect.

**Unnatural Language.** The response from the model is annotated as unnatural if it is phrased in a way that a native speaker would not typically use. This includes instances where words sound like direct translations from English, phrases that sound unnecessarily formal, or when a different language is used to answer.

**Stereotypical.** This includes responses containing stereotypical answers about a target country/region. For example, providing the most common traditional food in the country/region as an answer to a completely unrelated question would be considered a stereotypical response.

**Partially correct.** The response is annotated as partially correct when the model's response contains multiple answers and at least one is completely inapplicable to the general population of the country/region.

**Refusal.** This category indicates where the model declines to provide an answer despite the annotators having determined that a valid answer exists.

**Nonsensical.** Nonsensical answers include hallucinations from the model or are completely incorrect by not answering the question properly (e.g., answering "soccer" for a question about a sport played without a ball).

**Different country's view.** A response is annotated under this category if the model includes answers from the viewpoint of a different country/region. For instance, it includes answers from neighboring countries or countries sharing a similar yet different culture.

### 4.3.2 Human Evaluation Result

The summary of the human evaluation result by each error category is shown in Table 9. Detailed analysis is included in the main text.

We also present a more detailed human analysis of the responses from GPT-4 for selected countries/regions in this section, focusing primarily on under-represented cultures. All responses from the model were generated in respective local languages, but we present them here in English for the readers' convenience.

**Algeria (Arabic).** Stereotypical responses from the model were predominantly observed in food-related questions. Nearly all such responses included *couscous*, a traditional North African dish, even when irrelevant to the question. For example, the model suggested *couscous* and *baklava* as common picnic foods in Algeria, which is both inaccurate and somehow stereotypical.

Hallucinations were frequently encountered in responses to questions about celebrations or sports not commonly observed in Algeria. For instance, when asked about Halloween, the model referenced an unrelated old tradition and included the name of an equally unrelated sweet in Latin script, as shown below:

البوقالات (Boussou La Tmessou).

Another issue with the model's responses was the tendency to provide answers applicable to other Arabic-speaking countries, particularly Middle Eastern ones. This often led to culturally inaccurate or inappropriate responses for the Algerian context. For instance, when asked about the least favorite vegetable, the model mentioned *bamiya/bamieh*, the Middle Eastern name for okra. In Algeria, okra is called differently *(mloukhiya)* and is not commonly consumed nationwide. A similar misalignment with the Middle Eastern view was found in responses about local café brands and popular YouTube channels.

**Assam (Assamese).** The responses of the model often pointed towards Bihu, a cultural celebration of the Assamese people, even though it did not fit the context. It answered many questions with references to Bihu or Bihu-related activities. For instance, the model answered many food-related questions with *Pitha*, a traditional food item only served on special occasions like Bihu. The model also hallucinated by naming the most popular sports tournament in Assam as the *Bihu Tournament*, despite no such tournament existing in Assam.

**Azerbaijan (Azerbaijani).** The model often gave stereotypical answers related to traditional Azerbaijani dishes, irrespective of context. For example, it offered traditional foods as answers like *Qutab* and *Kebab* even for settings like amusement parks or fast food preferences, which are not the most typical or relevant choices in those contexts.

Additionally, the model often provided answers broadly applicable to people from post-Soviet or Eastern European regions rather than offering responses that uniquely represent Azerbaijan. Though these responses are not necessarily incorrect, they can be interpreted as lacking specificity. For instance, the model answered that the most famous leisure activity among retired men in Azerbaijan is *chess*, which is a viable option but is still more famous in Russia and Türkiye.

**China (Chinese).** The responses from the model were generally acceptable, with a few cases either stereotypical or biased toward Western culture. For instance, the model answered that the most famous sport played without a ball in China is *table tennis*, which is both stereotypical and nonsensical. It also answered that the most popular sports-related TV program in China is *Sports Scene*, a Chinese TV program broadcasted in English.

At certain times, the model demonstrated impressive capabilities, indicating its high cultural understanding of China. For instance, when answering questions related to Ramadan, the model showed a good understanding of the minor population in China. Though Ramadan is generally not observed in most parts of China, it is often observed in certain regions, particularly in the Ningxia province. The text below is the response from GPT-4 on the question, 'What do people from China eat in Ramadan?' (translated in English).

> Ramadan is the Islamic fasting month, mainly observed by Muslims. Chinese Muslims eat

> prepared food before sunrise (called “Suhur”) and break their fast (called “Iftar”) after sunset during Ramadan. They usually eat light, nutritious food, including fruits, vegetables, meat, beans, dairy products, and grains. Non-Muslim Chinese people do not eat any different food during Ramadan than usual.

**Ethiopia (Amharic).** Nonsensical answers were significantly prevalent, where the model often repeated the question itself as an answer. There were even answers containing typographic errors. Additionally, there were several cases where the model gave long texts of repeated words and phrases. Such incidents indicate the model’s limited ability to understand and use Amharic.

The model often gave answers commonly associated with Ethiopia but did not necessarily answer the question correctly. For instance, the model gave *Injera* as the answer for most of the food-related questions, possibly because ‘Injera’ is a well-known food item in Ethiopia. These answers were often regarded as stereotypical or even nonsensical.

**Greece (Greek).** Stereotypical answers were mostly from food-related questions, where the model gave a typical Greek dish as an answer to an irrelevant question. For instance, the model answered that the most popular flavor of crisps/chips is *feta cheese*, which is not a very popular choice among people.

There were also several instances where the model displayed biases towards the English culture. For example, it incorrectly stated that people in Greece eat *pumpkin pie* during Halloween, even though Halloween is not widely celebrated in Greece. It also answered that one of the most popular sports among elderly people is *golf*, a sport that is not as popular as other countries around the Mediterranean.

**Indonesia (Indonesian).** Most of the stereotypical answers came from the food category questions. The most popular choice from the model was *nasi goreng (fried rice)*, where the model even gave that as an answer to a question about the most popular wheat-based food item. Hallucinations were also common for questions requiring a person’s name, where the model provided the name of a completely unrelated person.

Though it was very rare, there were instances where the answers could be considered offensive, especially for questions related to religion. For example, the model incorrectly identified *Ketupat*, a dish commonly served during Muslim festivals in Indonesia, as the most common food served during Easter. Such answers may inadequately represent the Christian population in Indonesia.

An interesting example related to ‘different country’s view’ came from the following question: ‘What is installed in front of the house when a family member dies in your country?’. The model’s answer was *flying the flag at half mast*, a practice common in other countries during national mourning. However, this practice is not applicable when a family member dies in Indonesia. In Indonesia, people usually put up a yellow flag to indicate that someone has died in that area. There were many other instances where the model answered from the perspective of a different country. For example, it provided *Independence Day* as an answer to a question about the day of the year dedicated to fireworks in Indonesia. In Indonesia, people do not celebrate Independence Day by using fireworks.

**Iran (Persian).** Hallucinations were very common when answering questions that required a person’s name. For instance, it incorrectly identified the Mayor of Tehran as the most famous boxer, provided the coach’s name instead of the athlete’s, and even provided non-existent names.

In many cases, the model refused to answer because the question was considered illegal according to local laws. For instance, when asked about the most common alcoholic drink, the model responded that these drinks are illegal in Iran and, therefore, it could not provide an answer.

The model almost always provided answers to questions about a specific date based on the Gregorian calendar, even though people in Iran use the Solar Hijri calendar. While the answers were mostly correct when converted, the fact that both the questions and answers were in Persian suggests that the responses lacked cultural sensitivity.

**North Korea (Korean).** Offensive responses were heavily prevalent in North Korea, where the model answered *Kim Jong Un*, the current supreme leader of North Korea, for completely unrelated questions, such as the most popular fruit in North Korea or the type of shoes students wear at school.

Moreover, the responses from the model were biased towards the people from Pyongyang, the capital of North Korea. This phenomenon may stem from insufficient information about people from other areas in North Korea.

Another interesting finding was that the responses from the model were often phrased in the words used exclusively in South Korea. For instance, the answer given by the model for many food-related questions was *__n__aengmyeon (냉면)*, despite the fact that it is spelled differently in North Korea (*__r__aengmyon (랭면)*).

**South Korea (Korean).** Most incorrect responses that reflected the viewpoint of the other country were mainly due to the different age system used in South Korea. For instance, the model answered *19* for the question about the average age at which people go to university, whereas the most plausible answer would be '20' according to the South Korean age system. Such responses are surprising, as we have explicitly prompted the model to provide the answer using South Korea's traditional age-counting custom.

One interesting case was the question about the most famous family in South Korea. The model answered *Admiral Yi Sun-sin's family*, referencing a national hero who is very famous among people from South Korea, but not his family. Similarly, there were several instances where the model hallucinated by giving inaccurate answers tied to South Korea's traditional culture or history.

**West Java (Sundanese).** Unlike prior expectations that the model would wrongly provide answers applicable to people from all parts of Indonesia, as West Java is a specific region within the Indonesian country, the model tended to offer specific answers related to West Java. However, the problem was that these answers did not include a full understanding of the context. For instance, the model answered *Dodol Garut*, a traditional dessert from West Java, for a question asking about the food associated with Valentine's Day. Such a response is very stereotypical, considering that people in West Java also exchange chocolate for Valentine's Day, similar to other countries.

There were also errors in the language used by the model, where it answered in Indonesian instead of Sundanese.

## References

[1] Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*, 2024. URL `https://arxiv.org/abs/2405.15032`.

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. URL `https://arxiv.org/abs/2309.16609`.

[3] Andrew Bimba, Norisma Idris, Norazlina Khamis, and Nurul Noor. Stemming hausa text: using affix-stripping rules and reference look-up. *Language Resources and Evaluation*, 50, 07 2015. doi: 10.1007/s10579-015-9311-x.

[4] Kyle P. Johnson, Patrick Burns, John Stewart, and Todd Cook. Cltk: The classical language toolkit, 2014–2021. URL `https://github.com/cltk/cltk`.

[5] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL `https://aclanthology.org/2020.acl-main.560`.

[6] Anoop Kunchukuttan. The IndicNLP Library. `https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf`, 2020.

[7] Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. Seallms - large language models for southeast asia. *arXiv preprint arXiv:2312.00738*, 2023. URL `https://arxiv.org/abs/2312.00738`.

[8] Irwan Setiawan and Hung-Yu Kao. Sustem: An improved rule-based sundanese stemmer. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, apr 2024. ISSN 2375-4699. doi: 10.1145/3656342. URL `https://doi.org/10.1145/3656342`. Just Accepted.

[9] Kang Min Yoo, Jaegeun Han, Sookyo In, Heewon Jeon, Jisu Jeong, Jaewook Kang, Hyunwook Kim, Kyung-Min Kim, Munhyong Kim, Sungju Kim, Donghyun Kwak, Hanock Kwak, Se Jung Kwon, Bado Lee, Dongsoo Lee, Gichang Lee, Jooho Lee, Baeseong Park, Seongjin Shin, Joonsang Yu, Seolki Baek, Sumin Byeon, Eungsup Cho, Dooseok Choe, Jeesung Han, Youngkyun Jin, Hyein Jun, Jaeseung Jung, Chanwoong Kim, Jinhong Kim, Jinuk Kim, Dokyeong Lee, Dongwook Park, Jeong Min Sohn, Sujung Han, Jiae Heo, Sungju Hong, Mina Jeon, Hyunhoon Jung, Jungeun Jung, Wangkyo Jung, Chungjoon Kim, Hyeri Kim, Jonghyun Kim, Min Young Kim, Soeun Lee, Joonhee Park, Jieun Shin, Sojin Yang, Jungsoon Yoon, Hwaran Lee, Sanghwan Bae, Jeehwan Cha, Karl Gylleus, Donghoon Ham, Mihak Hong, Youngki Hong, Yunki Hong, Dahyun Jang, Hyojun Jeon, Yujin Jeon, Yeji Jeong, Myunggeun Ji, Yeguk Jin, Chansong Jo, Shinyoung Joo, Seunghwan Jung, Adrian Jungmyung Kim, Byoung Hoon Kim, Hyomin Kim, Jungwhan Kim, Minkyoung Kim, Minseung Kim, Sungdong Kim, Yonghee Kim, Youngjun Kim, Youngkwan Kim, Donghyeon Ko, Dughyun Lee, Ha Young Lee, Jaehong Lee, Jieun Lee, Jonghyun Lee, Jongjin Lee, Min Young Lee, Yehbin Lee, Taehong Min, Yuri Min, Kiyoon Moon, Hyangnam Oh, Jaesun Park, Kyuyon Park, Younghun Park, Hanbae Seo, Seunghyun Seo, Mihyun Sim, Gyubin Son, Matt Yeo, Kyung Hoon Yeom, Wonjoon Yoo, Myungin You, Doheon Ahn, Homin Ahn, Joohee Ahn, Seongmin Ahn, Chanwoo An, Hyeryun An, Junho An, Sang-Min An, Boram Byun, Eunbin Byun, Jongho Cha, Minji Chang, Seunggyu Chang, Haesong Cho, Youngdo Cho, Dalnim Choi, Daseul Choi, Hyoseok Choi, Minseong Choi, Sangho Choi, Seongjae Choi, Wooyong Choi, Sewhan Chun, Dong Young Go, Chiheon Ham, Danbi Han, Jaemin Han, Moonyoung Hong, Sung Bum Hong, Dong-Hyun Hwang, Seongchan Hwang, Jinbae Im, Hyuk Jin Jang, Jaehyung Jang, Jaeni Jang, Sihyeon Jang, Sungwon Jang, Joonha Jeon, Daun Jeong, Joonhyun Jeong, Kyeongseok Jeong, Mini Jeong, Sol Jin, Hanbyeol Jo, Hanju Jo, Minjung Jo, Chaeyoon Jung, Hyungsik Jung, Jaeuk Jung, Ju Hwan Jung, Kwangsun Jung, Seungjae Jung, Soonwon Ka, Donghan Kang, Soyoung Kang, Taeho Kil, Areum Kim, Beomyoung Kim, Byeongwook Kim, Daehee Kim, Dong-Gyun Kim, Donggook Kim, Donghyun Kim, Euna Kim, Eunchul Kim, Geewook Kim, Gyu Ri Kim, Hanbyul Kim, Heesu Kim, Isaac Kim, Jeonghoon Kim, Jihye Kim, Joonghoon Kim, Minjae Kim, Minsub Kim, Pil Hwan Kim, Sammy Kim, Seokhun Kim, Seonghyeon Kim, Soojin Kim, Soong Kim, Soyoon Kim, Sunyoung Kim, Taeho Kim, Wonho Kim, Yoonsik Kim, You Jin Kim, Yuri Kim, Beomseok Kwon, Ohsung Kwon, Yoo-Hwan Kwon, Anna Lee, Byungwook Lee, Changho Lee, Daun Lee, Dongjae Lee, Ha-Ram Lee, Hodong Lee, Hwiyeong Lee, Hyunmi Lee, Injae Lee, Jaeung Lee, Jeongsang Lee, Jisoo Lee, Jongsoo Lee, Joongjae Lee, Juhan Lee, Jung Hyun Lee, Junghoon Lee, Junwoo Lee, Se Yun Lee, Sujin Lee, Sungjae Lee, Sungwoo Lee, Wonjae Lee, Zoo Hyun Lee, Jong Kun Lim, Kun Lim, Taemin Lim, Nuri Na, Jeongyeon Nam, Kyeong-Min Nam, Yeonseog Noh, Biro Oh, Jung-Sik Oh, Solgil Oh, Yeontaek Oh, Boyoun Park, Cheonbok Park, Dongju Park, Hyeonjin Park, Hyun Tae Park, Hyunjung Park, Jihye Park, Jooseok Park, Junghwan Park, Jungsoo

Park, Miru Park, Sang Hee Park, Seunghyun Park, Soyoung Park, Taerim Park, Wonkyeong Park, Hyunjoon Ryu, Jeonghun Ryu, Nahyeon Ryu, Soonshin Seo, Suk Min Seo, Yoonjeong Shim, Kyuyong Shin, Wonkwang Shin, Hyun Sim, Woongseob Sim, Hyejin Soh, Bokyong Son, Hyunjun Son, Seulah Son, Chi-Yun Song, Chiyoung Song, Ka Yeon Song, Minchul Song, Seungmin Song, Jisung Wang, Yonggoo Yeo, Myeong Yeon Yi, Moon Bin Yim, Taehwan Yoo, Youngjoon Yoo, Sungmin Yoon, Young Jin Yoon, Hangyeol Yu, Ui Seon Yu, Xingdong Zuo, Jeongin Bae, Joungeun Bae, Hyunsoo Cho, Seonghyun Cho, Yongjin Cho, Taekyoon Choi, Yera Choi, Jiwan Chung, Zhenghui Han, Byeongho Heo, Euisuk Hong, Taebaek Hwang, Seonyeol Im, Sumin Jegal, Sumin Jeon, Yelim Jeong, Yonghyun Jeong, Can Jiang, Juyong Jiang, Jiho Jin, Ara Jo, Younghyun Jo, Hoyoun Jung, Juyoung Jung, Seunghyeong Kang, Dae Hee Kim, Ginam Kim, Hangyeol Kim, Heeseung Kim, Hyojin Kim, Hyojun Kim, Hyun-Ah Kim, Jeehye Kim, Jin-Hwa Kim, Jiseon Kim, Jonghak Kim, Jung Yoon Kim, Rak Yeong Kim, Seongjin Kim, Seoyoon Kim, Sewon Kim, Sooyoung Kim, Sukyoung Kim, Taeyong Kim, Naeun Ko, Bonseung Koo, Heeyoung Kwak, Haena Kwon, Youngjin Kwon, Boram Lee, Bruce W. Lee, Dagyeong Lee, Erin Lee, Euijin Lee, Ha Gyeong Lee, Hyojin Lee, Hyunjeong Lee, Jeeyoon Lee, Jeonghyun Lee, Jongheok Lee, Joonhyung Lee, Junhyuk Lee, Mingu Lee, Nayeon Lee, Sangkyu Lee, Se Young Lee, Seulgi Lee, Seung Jin Lee, Suhyeon Lee, Yeonjae Lee, Yesol Lee, Youngbeom Lee, Yujin Lee, Shaodong Li, Tianyu Liu, Seong-Eun Moon, Taehong Moon, Max-Lasse Nihlenramstroem, Wonseok Oh, Yuri Oh, Hongbeen Park, Hyekyung Park, Jaeho Park, Nohil Park, Sangjin Park, Jiwon Ryu, Miru Ryu, Simo Ryu, Ahreum Seo, Hee Seo, Kangdeok Seo, Jamin Shin, Seungyoun Shin, Heetae Sin, Jiangping Wang, Lei Wang, Ning Xiang, Longxiang Xiao, Jing Xu, Seonyeong Yi, Haanju Yoo, Haneul Yoo, Hwanhee Yoo, Liang Yu, Youngjae Yu, Weijie Yuan, Bo Zeng, Qian Zhou, Kyunghyun Cho, Jung-Woo Ha, Joonsuk Park, Jihyun Hwang, Hyoung Jo Kwon, Soonyong Kwon, Jungyeon Lee, Seungho Lee, Seonghyeon Lim, Hyunkyung Noh, Seungho Choi, Sang-Woo Lee, Jung Hwa Lim, and Nako Sung. Hyperclova x technical report. *arXiv preprint arXiv:2404.01954*, 2024. URL `https://arxiv.org/abs/2404.01954`.

[10] Taha Zerrouki. qalsadi, arabic mophological analyzer library for python., 2012. URL `https://pypi.python.org/pypi/qalsadi`.

[11] Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*, 2024. URL `https://arxiv.org/abs/2402.07827`.