wangkongqiang at SemEval-2026 Task 7: Everyday Knowledge Across Diverse Languages and Cultures

Kongqiang Wang[1], **Peng Zhang**[2], **Qingli Tan**[3]
[1,2]School of Information Science and Engineering, Yunnan University,
[3]School of College of Ecology and Environment, Yunnan University,
Kunming 650500, Yunnan, China
[1]wangkongqiang60@gmail.com, [2]zpp1219@gmail.com, [3]tanqingli@stu.ynu.edu.cn

Abstract

This paper presents our system developed for the SemEval-2026 Task 7: Everyday Knowledge Across Diverse Languages and Cultures on Subtask 1: Short Answer Questions (SAQ), and Subtask 2: Multiple-Choice Questions (MCQ). To this end, we focus on models' cultural competence across 26 languages and 30 countries using four different versions large language models (LLMs): deepseek-v3.2-exp, qwen-max, qwen-plus, and qwen3-next-80b-a3b-instruct. We experiment with (1) the trial and test dataset is analyzed visually, (2) use the large language generative model to perform generate or select the answer that it deems correct on the trial and test dataset through prompts, and (3) many prompt engineering approaches of generative models are evaluated on the trial dataset. We further study the influence of different hyperparameters on the generative model and select the best single model for the prediction of the test dataset. Our submission achieved the good ranking place in the test dataset leaderboard. For Subtask 1 (SAQ), the evaluation criteria for this task mainly consist of the aggregate results of the 23 languages: ar-EG, ar-MA, ar-SA, bg-BG, el-GR, en-AU, and so on, and they are measured using the accuracy score. For Subtask 2 (MCQ), this task is essentially a multiple-choice task for questions text. Performance will be evaluated using accuracy score. In other words, this subtask evaluated using accuracy score based on the correctness of the selected answer across different languages and cultural contexts. For Subtask 1 (SAQ) and Subtask 2 (MCQ), our best approach is to obtain the results in test dataset are accuracy score 51.4689 and accuracy score 80.26 separately. For the final ranking, organizers will use the aggregate results of accuracy score. Even so, our approach has yielded good results.

1 Introduction

The official organizers of SemEval (Ghosh et al., 2026) held SemEval 2026 - Everyday Knowledge Across Diverse Languages and Cultures in Task 7 (Ousidhoum et al., 2026) in the first half of 2026. The main content of this task is models' understanding of everyday knowledge in diverse multilingual and multicultural contexts. The purpose of this task and the ideal goal to be achieved are: To understand cultural awareness in language models. In the other words, understand cultural awareness thinking at its roots. This is a shared task on evaluating cultural awareness in language models across 26 languages and 30 countries or regions. Participants will use an extended BLEnD benchmark to test models' understanding of everyday knowledge in diverse multilingual and multicultural contexts.

Online multicultural opinion is the colorful and diverse between social, political, or identity groups. The opinion of multilingual thinking on the Internet has become a growing concern, as it often precedes life speech, everyday discourse, and social development. Based on this background, we now proceed to introduce the overview of this sharing task. The global deployment of large language models (LLMs) necessitates cultural awareness and competence. However, LLMs often exhibit a significant deficiency in culture-specific knowledge, particularly concerning under-resourced languages and regions. They tend to generate responses that reflect Western-centric perspectives or the stereotypes present in their training data. Given that existing benchmarks predominantly rely on monolingual datasets or online resources like Wikipedia, which often fail to capture the nuanced realities of everyday life across diverse cultural contexts, the official team has developed BLEnD, a benchmark that comprehensively evaluates LLMs' understanding of everyday knowledge in multilingual and multicultural contexts. In this shared task, they will provide an extended version of BLEnD, enabling the evaluation of language models' cultural competence across 26 languages and 30 countries. For the first time, Semeval organizers introduce

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 247–255

a Knowledge Q&A task: Everyday Knowledge Across Diverse Languages and Cultures, aimed at develop the cultural awareness and competence of large language models (LLMs). The task focuses on the study of multicultural and multievent everyday knowledge, capturing the complexity of online discourse across diverse contexts. Participants may participate in one or more of the following two sub-tasks: Short Answer Questions (SAQ), and Multiple-Choice Questions (MCQ). To be more specific, we are encouraging to join them in tackling two synergistic subtasks. **Track 1: Short Answer Questions (SAQ)**. Participants will test their models on short-answer questions (SAQs) to ensure they can accurately generate responses while accounting for cultural and linguistic diversity. This track will include 26 languages. Given questions in a given language, responses will be tested in that language, and correctness will be determined based on alignment with human-annotated answers from BLEnD. **Track 2: Multiple-Choice Questions (MCQ)**. In this track, questions are provided in English only. Each question will come with four answer options, each representing a cultural perspective from a different country or region, i.e., the one that received the highest number of votes for a given country. To ensure fairness, questions are filtered to exclude those marked as culturally irrelevant or unclear by human annotators. Each multiple-choice question includes four answer options, with no more than one option representing any of the other countries or regions. The developed model is assessed based on its ability to identify the culturally appropriate choice for each question per region. Based on the predict and choice task background of everyday knowledge question text, we propose the prompt words guidance for generation method based on qwen and deepseek series large language models (LLMs) for generating answer content or selecting the answer it think is correct.

We developed for the SemEval-2026 Task 7: Everyday Knowledge Across Diverse Languages and Cultures. on Subtask 1: Short Answer Questions (SAQ). Evaluation will use a question-and-answer dataset containing languages from multiple regions. The following provides a detailed explanation of how many languages are used for testing the cultural perception ability of large language models (LLMs). The languages/regions included in our shared task, see Table 1.

on Subtask 2: Multiple-Choice Questions (MCQ). In this track, questions in trial dataset aren't provided in English only. The authorities have made efforts to enable the large language model (LLM) to have the ability to understand multiple languages question options. In the trial dataset, the questions and options presented are in multiple languages. However, questions in test dataset are provied in English only for the regions listed in Table 2.

Semeval organizers will evaluate each submission using accuracy based on the alignment of the generated answer with human annotations. Notably, their evaluation accounts for variations in responses, which ensures a more robust assessment. Specifically, in the SAQ track, a model-generated answer is marked as correct if it matches any of the

Table 1: The languages/regions included for Subtask 1: Short Answer Questions (SAQ) are described.

Code	Language (Region)
am-ET	Amharic (Ethiopia)
ar-DZ	Arabic (Algeria)
ar-EG	Arabic (Egypt)
ar-MA	Arabic (Morocco)
ar-SA	Arabic (Saudi Arabia)
as-AS	Assamese (Assam, India)
az-AZ	Azerbaijani (Azerbaijan)
bg-BG	Bulgarian (Bulgaria)
el-GR	Greek (Greece)
en-AU	English (Australia)
en-GB	English (United Kingdom)
en-SG	English (Singapore)
en-US	English (United States)
es-EC	Spanish (Ecuador)
es-ES	Spanish (Spain)
es-MX	Spanish (Mexico)
eu-PV	Basque (Basque Country, Spain)
fa-IR	Persian/Farsi (Iran)
fr-FR	French (France)
ga-IE	Irish (Ireland)
ha-NG	Hausa (Northern Nigeria)
id-ID	Indonesian (Indonesia)
ja-JP	Japanese (Japan)
ko-KP	Korean (North Korea)
ko-KR	Korean (South Korea)
ms-SG	Malay (Singapore)
su-JB	Sundanese (West Java, Indonesia)
sv-SE	Swedish (Sweden)
ta-LK	Tamil (Sri Lanka)
ta-SG	Tamil (Singapore)
tl-PH	Tagalog (Philippines)
zh-CN	Chinese (China)
zh-SG	Singaporean Mandarin (Singapore)
zh-TW	Taiwanese Mandarin (Taiwan)
en-AS	English (Assam)
en-AZ	English (Azerbaijan)
en-BG	English (Bulgaria)
en-CN	English (China)
en-DZ	English (Algeria)
en-EC	English (Ecuador)
en-EG	English (Egypt)
en-ES	English (Spain)
en-ET	English (Ethiopia)
en-FR	English (France)
en-GR	English (Greece)
en-ID	English (Indonesia)
en-IE	English (Ireland)
en-IR	English (Iran)
en-JB	English (West Java, Indonesia)
en-JP	English (Japan)
en-KP	English (North Korea)
en-KR	English (South Korea)
en-LK	English (Sri Lanka)
en-MA	English (Morocco)
en-MX	English (Mexico)
en-NG	English (Nigeria)
en-PH	English (Philippines)
en-PV	English (Basque Country, Spain)
en-SA	English (Saudi Arabia)
en-SE	English (Sweden)
en-TW	English (Taiwan)

Table 2: The regions/countries included for Subtask 2: Multiple-Choice Questions (MCQ) are described.

Code	Region (Country)
am-ET	Ethiopia
ar-DZ	Algeria
ar-EG	Egypt
ar-MA	Morocco
ar-SA	Saudi Arabia
as-AS	Assam (India)
az-AZ	Azerbaijan
bg-BG	Bulgaria
el-GR	Greece
en-AU	Australia
en-GB	United Kingdom
en-US	United States
es-EC	Ecuador
es-ES	Spain
es-MX	Mexico
eu-PV	Basque Country, Spain
fa-IR	Iran
fr-FR	France
ga-IE	Ireland
ha-NG	Nigeria
id-ID	Indonesia
ja-JP	Japan
ko-KP	North Korea
ko-KR	South Korea
su-JB	West Java, Indonesia
sv-SE	Sweden
ta-LK	Sri Lanka
tl-PH	Philippines
zh-CN	China
zh-SG	Singapore

responses provided by human annotators for the same question, and in the MCQ track, accuracy is calculated based on the correctness of the selected answer. More details about the evaluation protocol can be found in Myung et al. (Myung et al., 2024) conducted relevant research in 2024. The code of our method is available on our GitHub website[1].

2 Related Work

SemEval in previous years has introduced tasks focusing on LLM capabilities and cultural perception ability (Vazquez et al., 2025; Ramakrishna et al., 2025; D'souza et al., 2025; Brekhof et al., 2024) to evaluate internal potential elements and potential content of the large language model (LLM). These tasks provided chances with using LLMs and fully leverage their capabilities, which have been extensively utilized for content generation tasks and knowledge question-answering tasks.

2.1 Prompt Engineering

Prompt engineering refers to a methodology and technical system that carefully designs input prompts to guide the models to generate more accurate, stable, and task-compliant outputs when using large language models (LLMs). Unlike traditional machine learning that relies on large-scale labeled data and parameter fine-tuning (Muhammad et al., 2025b), prompt engineering emphasizes controlling the model with language, significantly improving the model's performance in specific tasks under low-cost and low-threshold conditions. With the wide application of general large models in fields such as natural language processing (NLP), information extraction, code generation, and scientific research, prompt engineering has become a key link connecting the capabilities of models with practical application scenarios.

Large language models (LLMs) usually rely on autoregressive mechanisms to predict the next most likely word or symbol based on the existing context. The prompt essentially constitutes the context environment for model reasoning. Its content, structure and expression method will directly affect the generation path and final result of the model. From a cognitive perspective, the core of prompt engineering lies in:

- **Clarify Task Boundaries**: Tell the model what to do and what not to do;
- **Reduce Ambiguity Space**: Lower the uncertainty brought about by the free play of the model through structured description;
- **Activate Potential Capabilities**: Guide the model to invoke the relevant knowledge and patterns it has learned during the pre-training stage.

The following briefly describes the basic components of prompt. A high-quality prompt usually consists of the following elements:

- **Task Instruction**: Clearly state the specific tasks that the model needs to complete, such as classification, summarization, extraction or generation.
- **Background Information**: Provide the necessary context or domain background for the model to narrow down the scope of reasoning.
- **Input Data**: Text, data or the problem itself that needs to be processed by the model.
- **Output Constraints**: Impose restrictions on the output form, length, language or structure, such as "Return the result in JSON format".

In practical applications, not all prompts need to fully incorporate the above elements. They should be flexibly adjusted according to the complexity of the task.

The following briefly describes the common prompt engineering strategies.

- **Zero-shot Prompting**: Zero-shot prompting means to make a model complete a task directly by giving instructions without providing examples. This approach relies on the general knowledge and reasoning ability of the model and is suitable for scenarios with clear task definitions and low difficulty.
- **Few-shot Prompting**: Few-shot prompting helps the model understand the task pattern by providing a few examples (input-output pairs) in the prompt. This method can significantly enhance the model's performance in structured tasks and specialized domain tasks.
- **Chain-of-Thought Prompting**: Chain-of-Thought (CoT) prompts to explicitly generate intermediate reasoning processes by guiding

[1] https://github.com/WangKongQiang/Semeval2026-Task7

the model to think step by step, thereby enhancing the accuracy of complex reasoning tasks such as mathematical calculations and logical inferences.

- **Role-based Prompting**: By setting roles for the model such as "You are an environmental science researcher", it can guide the model to adopt a specific perspective, language style and professional depth, which is often used in academic writing and domain analysis.

2.2 Qwen Large Language Model (LLM)

Qwen (Tongyi Qianwen) large language model (LLM) is a series of general-purpose large language models (LLMs) launched by Alibaba DaMo academy, aiming to provide high-performance, scalable and open language understanding and generation capabilities. The qwen series of models cover multiple versions ranging from lightweight to large parameter scales, and can be adapted to various scenarios such as scientific research experiments, engineering deployments, and industrial applications. In terms of overall design, the qwen model takes the Transformer architecture as its core, inheriting and expanding the technical routes of current mainstream autoregressive language models. It demonstrates strong comprehensive capabilities in multilingual tasks (Muhammad et al., 2025a) such as Chinese and English, especially showing good stability in instruction understanding, text generation, and complex reasoning tasks.

The following elaborates on the model architecture and technical features in terms of the three aspects: infrastructure and position encoding, as well as context modeling. The qwen series models are based on the Decoder-only Transformer architecture and model the context of long texts through a multi-layer Self-Attention mechanism. During the pre-training stage, the model uses large-scale corpora for unsupervised learning, enabling it to capture the statistical features and semantic structure of the language. Compared with traditional recurrent neural network models, the Transformer architecture has significant advantages in parallel computing capabilities and long-distance dependency modeling, providing a technical foundation for large-scale language modeling. To enhance the model's modeling ability for long sequences, the qwen model has been optimized in terms of position encoding strategy and attention mechanism design, enabling the model to maintain stable semantic consistency under longer context conditions. This feature is particularly important in tasks such as understanding long documents, multi-round conversations, and executing complex instructions.

In previous studies (Belay et al., 2025), prompt engineering presents several advantages. The prompt engineering approach can reduce the errors from insufficient examples by expanding the learning of the text context or can make the system generation more robust. In our study, using the qwen series and deepseek-v3.2-exp generative model to perform prompt engineering on the trial and test dataset through prompts to generate large amounts of correct answers while making use of information from the prompts and questions. Previous research has demonstrated that appropriate prompt engineering can achieve remarkable success.

In our study, we aim to use multiple large language models (LLMs) to assess semantic answers. When models are invoked on diverse datasets with different prompts, they may produce varied predictions on semantic answers, and prompt engineering is not a one-off task but a continuous iterative process. Common optimization ideas include:

- **Comparative Experiment**: Compare the output quality of different prompts on the same task;
- **Error Analysis**: Summarize the types of output errors of the model and adjust the prompt accordingly;
- **Constraint Reinforcement**: Reduce irrelevant or erroneous output through more explicit formats and rules;
- **Automated Search**: Systematically optimize prompts in combination with scripts or algorithms.

Evaluation metrics are usually set based on specific tasks, such as accuracy, consistency, interpretability or manual scoring results. For both of our track tasks, they are based on the accuracy metric that measures consistency with the manually annotated answers. We use large language models (LLMs) mainly from the following models: deepseek-v3.2-exp, qwen-max, qwen-plus, and qwen3-next-80b-a3b-instruct. These models can be

invoked to be controlled by the large model service platform Bailian of Alibaba cloud[2].

3 Methodology

3.1 Overall Architecture

For Subtask 1 (SAQ), we respectively used the following four different versions of large language models (LLMs) as the backbone system model to perform answer generation: $deepseek-v3.2-exp$, $qwen-max$, $qwen-plus$, and $qwen3-next-80b-a3b-instruct$. This method using simplified prompts mainly involves constraints on the generation of answer content. Finally, we aggregate the generated answer content results for each question sample and save the final aggregated predictions in tsv format.

For Subtask 2 (MCQ), the pursued approach involves using a multiple-choice questions selection system composed of the qwen series and $deepseek-v3.2-exp$ generative model and then generate the option that the large language model (LLM) considers to be the correct choice based on the prompts. The overall experimental procedure is shown in Figure 1.

3.2 Implementation Step

For Subtask 1: Short Answer Questions (SAQ), our experiment mainly involes these steps:

First, register on the Alibaba cloud Bailian platform to obtain the API key that can be used to call the model square service. The large language model (LLM) can be used from the official website of large language model (LLM) service platform - Bailian control console[3].

Second, configure the system environment and install the OpenAI dependencies. By calling the corresponding large language models (LLMs) on the large language model (LLM) service platform of Bailian console through the OpenAI dependency package.

Third, apply each online large language models (LLMs) to perform inference. Specifically, use simplified prompts to constraints on answers for this specific answer generation task. By aggregating the prediction results from all questions, the predicted minimal answers were obtained and saved in tsv format.

For Subtask 2: Multiple-Choice Questions (MCQ), our experiment mainly involves these steps:

First, set the API key, get from Alibaba cloud Bailian platform. The API key can be applied from the official website of large language model (LLM) service platform - Bailian control console.

Second, configure the system environment and install the OpenAI dependencies. By calling the corresponding large language models (LLMs) on the large language model (LLM) service platform of Bailian console through the OpenAI dependency package.

Third, since the content of the multiple_choice_options column provided by the trial dataset or test dataset is the information that the model requires as question options input content, they are selected by using the markers A,B,C, or D separately, and the question column content was retained as the raw question content. By invoking the API key of Alibaba cloud Bailian platform using the qwen series and $deepseek-v3.2-exp$ generative model to perform multiple-choice question answer selection on the trial dataset or test dataset through prompts to generate large amounts of marker texts, which is as the only correct answer selected for each question options.

Further, using the dictionary tool for conversion, the options labeled as A, B, C and D are transformed into one-hot format. Finally, all the one-hot contents of the selected options are saved in tsv format. Two type prompts for Subtask 1: Short Answer Questions (SAQ) and Subtask 2: Multiple-Choice Questions (MCQ) respectively, see Figure 2.

4 Results and Analysis

4.1 Trial Dataset Analysis

The question length and lang_region label of trial dataset are described in Table 3. The labels here are only analyzed on Subtask 1: Short Answer Questions (SAQ). Subtask 2: Multiple-Choice Questions (MCQ) the same as Subtask 1. The length and quantity distribution of trial question text data are analyzed in Figure 3. Distribution of the size of question texts for each language/region in Figure 4. It shows the number of percentages relative to each language class for various cases. Language classes here only analyze Subtask 1: Short Answer Questions(SAQ).

[2]https://www.aliyun.com/

[3]https://bailian.console.aliyun.com/

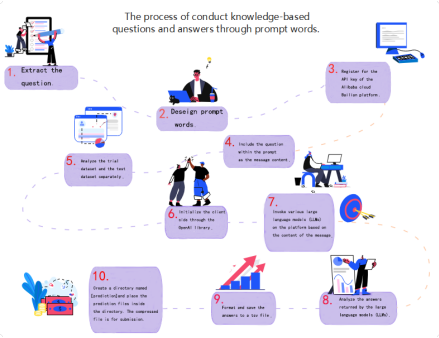

Figure 1: The process of conduct knowledge-based questions and answers through prompt words.

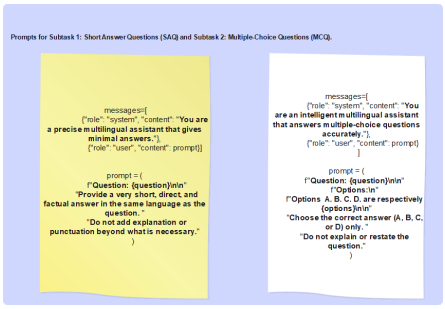

Figure 2: Prompts for Subtask 1: Short Answer Questions (SAQ) [*Left*] and Subtask 2: Multiple-Choice Questions (MCQ) [*Right*].

Table 3: The question text data situation and the number of lang_region labels for Subtask 1: Short Answer Questions (SAQ) are described.

Question Text Length	Value
count	148.000000
mean	9.351351
std	4.726341
min	1.000000
25%	6.000000
50%	9.000000
75%	12.000000
max	27.000000

Trial Dataset Lang_Region Label	Count
es-EC	8
tl-PH	8
fr-FR	8
ms-SG	7
ar-EG	7
bg-BG	7
ta-LK	7
ga-IE	7
eu-ES	7
en-AU	7
ar-SA	7
ar-MA	7
ja-JP	7
ta-SG	7
zh-SG	7
el-GR	5
ko-KR	5
id-ID	5
es-MX	5
es-ES	5
zh-CN	5
en-GB	5
fa-IR	5

4.2 Experimentation Configuration

For the sake of completeness and in an attempt to improve the results obtained by the large language models (LLMs). For $deepseek-v3.2-exp$, $qwen-max$, $qwen-plus$, and $qwen3-next-80b-a3b-instruct$, the four different hyperparameters were used respectively: See Table 4.

4.3 Trial Dataset Result

The following Table 5 records the official results of SemEval-2026 Task 7: Everyday Knowledge Across Diverse Languages and Cultures, on Subtask 1: Short Answer Questions (SAQ), and Subtask 2: Multiple-Choice Questions (MCQ). The metrics recorded by black bold text is the best (winning) approach in the evaluation task of the trial dataset for multiple languages.

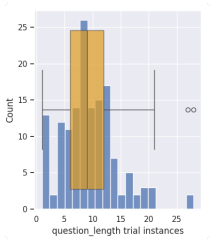

Figure 3: The length and quantity distribution of trial question text data are analyzed.

Table 4: Experimentation configuration hyperparameters for the four different large language models (LLMs).

model	Hyperparameter	Values
deepseek-v3.2-exp	maximum_input_length	96K
	maximum_output_length	64K
	context_length	128K
qwen-plus	maximum_input_length	997K
	maximum_output_length	32K
	maximum_input_length (thinking)	995K
	maximum_output_length (thinking)	32K
	context_length	1M
	maximum_length_of_the_thought_chain	80K
qwen-max	maximum_input_length	30K
	maximum_output_length	8K
	context_length	32K
qwen3-next-80b-a3b-instruct	maximum_input_length	126K
	maximum_output_length	32K
	context_length	128K

4.4 Test Dataset Result

The following Table 6 records the official results of SemEval-2026 Task 7: Everyday Knowledge Across Diverse Languages and Cultures, on Subtask 1: Short Answer Questions (SAQ), and Sub-

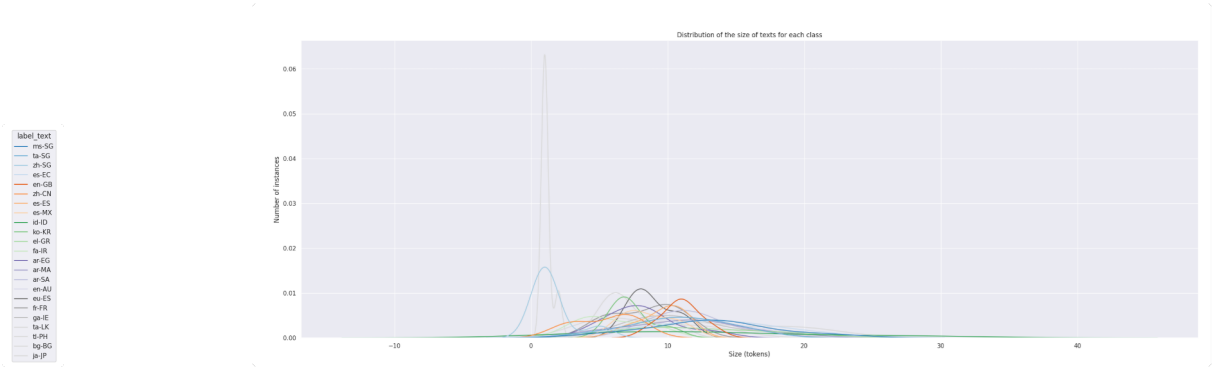

Figure 4: Distribution of the size of question texts for each language class. The horizontal axis represents the size of the problem text, while the vertical axis shows the percentage distribution of different language types questions across various length values.

Table 5: The trial dataset experiment situation detailed results are described.

Subtask	Language	Main Technologies	Accuracy
1	|overall|	deepseek-v3.2-exp	40.54
1	|overall|	**qwen-max**	**41.89**
1	|overall|	qwen-plus	31.08
1	|overall|	qwen3-next-80b-a3b-instruct	37.16
Subtask	**Language**	**Main Technologies**	**Accuracy**
2	|overall|	deepseek-v3.2-exp	87.16
2	|overall|	qwen-max	87.84
2	|overall|	**qwen-plus**	**91.89**
2	|overall|	qwen3-next-80b-a3b-instruct	89.19

Table 6: The test dataset experiment situation detailed results are described.

Subtask	Language	Main Technologies	Accuracy
1	|overall|	deepseek-v3.2-exp	-
1	|overall|	**qwen-max**	**51.4689**
1	|overall|	qwen-plus	-
1	|overall|	qwen3-next-80b-a3b-instruct	-
Subtask	**Language**	**Main Technologies**	**Accuracy**
2	|overall|	deepseek-v3.2-exp	-
2	|overall|	qwen-max	-
2	|overall|	**qwen-plus**	**80.26**
2	|overall|	qwen3-next-80b-a3b-instruct	-

task 2: Multiple-Choice Questions (MCQ). The metrics recorded by black bold text is the best (winning) approach in the evaluation task of the test dataset for multiple languages.

4.5 Question Text Words and Lang_Region Labels Biased Performance

From Figure 3 of the visual analysis, we can observe that 75% of question texts in trial dataset, either in the chart or in the previous model input column content, have no more than 15 words. This information could be useful in determining the size of the input tokens for large language models (LLMs), or when the size limit for the tokens consumed by the large language model (LLM) needs to be set.

SemEval-2026 Task 7: Everyday Knowledge Across Diverse Languages and Cultures, on Subtask 1: Short Answer Questions (SAQ). This task is essentially a text generation task, which requires the generate concise answers to the corresponding questions separately. Distribution of question characters length per trial data sample, as show in Figure 5. The developed model needs to understand various language questions and generate answers in the corresponding languages. The overall quantity statistics of each language in questions for test dataset, see Table 7 (*Upper*). A correct answer generation requires the results are in line with the objective facts and are consistent with the standard answers. In the test dataset, there are significant differences in the occurrence frequencies of the 22 types of annotation languages. Among them, the number of English language is the largest (15500 markers), indicating that most of the questions are in English. Arabic language comes second (2000 markers), reflecting the second most prominent issue is the cultural problem of Algeria. The number of Spainish and Chinese is similar (1500 markers), indicating that the annotation questions of Chinese and Spanish in the test dataset is relatively balanced. Tamil and Korean question texts respectively are 1000 markers. The number of other languages is the smallest (500 markers each language), indicating that the Japanese, Tagalog, Swedish and so on in some question texts are scarce.

on Subtask 2: Multiple-Choice Questions (MCQ). This task is Multiple-choice questions based on natural language understanding problems. Each question instance can have only one choice, and you need to predict which choice each question instance belongs to. In the specific cases (English language) we focus on, there may be up to four different selectable options: A, B, C and D. For this task, we will look at the quantity distribution followed by each question text lang_region, as shown in Table 7 (*Down*). In this case, percentages can be assessed because of the labels scattered. Based on the analysis of the valid multiple regions English language question in the test dataset, it is found

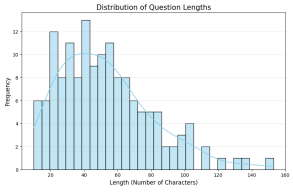

Figure 5: The question characters length distribution of trial data samples are analyzed.

Table 7: The overall quantity statistics of each type in languages for Subtask 1: Short Answer Questions (SAQ) and Subtask 2: Multiple-Choice Questions (MCQ) test dataset are described.

Subtask	Test Dataset Lang_Region Type	Count
1	en	15500
	ar	2000
	es	1500
	zh	1500
	ta	1000
	ko	1000
	ja	500
	tl	500
	sv	500
	su	500
	fa	500
	ms	500
	am	500
	ga	500
	ha	500
	el	500
	fr	500
	bg	500
	eu	500
	az	500
	as	500
	id	500
Subtask	**Test Dataset Lang_Region Type**	**Count**
2	es	4807
	ko	4697
	en	4622
	ar	3955
	fa	3699
	am	2863
	el	2734
	as	2451
	zh	2357
	su	2345
	az	2297
	ha	2008
	id	1995
	tl	1327
	ta	1114
	eu	1075
	ga	856
	bg	648
	sv	447
	ja	410
	fr	307

that approximately 46.3% of the question texts contain the culture related to Spanish, Korean, English, Arabic, Persian. while the question texts with other regions culture account for about 53.7%. Overall, the regions of qustions are roughly balanced.

5 Conclusion

Our system employs openai python package approach to invoke the online large language model (LLM) to address questions related to the culture of the specific region (Agirre et al., 2014), acquiring results from multiple large language models (LLMs): $deepseek - v3.2 - exp$, $qwen - max$, $qwen - plus$, and $qwen3 - next - 80b - a3b - instruct$. The hyperparameter is following: maximum_input_length is 30K, maximum_output_length is 8K, context_length is 32K. The dataset usage is shown in Table 8. Our findings suggest that questions and answers semantic relatedness can be deduced from a variety of sources. Although some questions of the culture in the relevant regions (e.g. Indonesian and Assamese region) may not perform strongly in online large language models (LLMs) specifically designed to general knowledge questions. The results demonstrate that these most questions make a response by online large language models (LLMs) can outperform many individual state-of-the-art systems and achieve a better correlation with human judgment on semantic relatedness (Siino, 2024) when used in the knowledge question-and-answer scenarios of each region.

Table 8: Use dataset supported by Semeval-2026 Task 7: Everyday Knowledge Across Diverse Languages and Cultures on Subtask 1: Short Answer Questions (SAQ) and Subtask 2: Multiple-Choice Questions (MCQ). The style is based on raw dataset.

Dataset Input	Description	Use or Not
extended version of BLEnD (Myung et al., 2024) dataset	dataset for evaluating cultural awareness in language models, it across 26 languages and 30 countries or regions.	yes
other dataset	use external or additional corpora.	no

6 Limitation and Future Work

Our experiments are only based on English language test datasets in Subtask 2: Multiple-Choice Questions (MCQ). Constrained by the use of the Alibaba cloud Bailian platform and the availability of online large language models (LLMs), it is regrettable that we did not offer insights into other Asian and African languages (Vaidya et al., 2024) for Subtask 2: Multiple-Choice Questions (MCQ). In future research, studies on low-resource languages will be valuable, including tasks such as data collection, annotation, and pre-training large language models (LLMs) tailored to these languages.

Acknowledgments

We are very grateful for the assistance and discussions provided by Semeval-2026 Task 7: Everyday Knowledge Across Diverse Languages and Cultures leaders and organizers.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, page 81–91, Dublin, Ireland. Association for Computational Linguistics.

Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Grigori Sidorov, Dietrich Klakow, Philip Slusallek, Olga Kolesnikova, and Seid Muhie Yimam. 2025. Evaluating the capabilities of large language models for multi-label emotion understanding. In

Proceedings of the 31st International Conference on Computational Linguistics, pages 3523–3540, Abu Dhabi, UAE. Association for Computational Linguistics.

Thijs Brekhof, Xuanyi Liu, Joris Ruitenbeek, Niels Top, and Yuwen Zhou. 2024. Groningen team D at SemEval-2024 task 8: Exploring data generation and a combined model for fine-tuning LLMs for multidomain machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 391–398, Mexico City, Mexico. Association for Computational Linguistics.

Jennifer D'souza, Sameer Sadruddin, Holger Israel, Mathias Begoin, and Diana Slawig. 2025. SemEval-2025 task 5: LLMs4Subjects - LLM-based automated subject tagging for a national technical library's open-access catalog. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2570–2583, Vienna, Austria. Association for Computational Linguistics.

Debanjan Ghosh, Kai North, Ekaterina Kochmar, Mamoru Komachi, and Marcos Zampieri, editors. 2026. *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics, San Diego, United States.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufino Cardenas, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Andrew Piper, Alexander Panchenko, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, and Saif M. Mohammad. 2025a. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *Preprint*, arXiv:2502.11926.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, Yi Zhou, and Saif M. Mohammad. 2025b. SemEval task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146.

Nedjma Ousidhoum, Junho Myung, Carla Perez-Almendros, Jiho Jin, Amr Keleg, Meriem Beloucif, Yi Zhou, Rodrigo Agerri, Vladimir Araujo, Naomi Baes, James Barry, Joanne Boisson, Nancy F. Chen, Christine de Kock, Aleksandra Edwards, Joseba Fernandez de Landa, Mohamed Fazli Imam, Huda Hakami, Shu-Kai Hsieh, Joseph Marvin Imperial, Roy Ka-Wei Lee, Chenyang Lyu, Younes Samih, Johan Sjons, Bryan Tan, Asahi Ushio, Weihua Zheng, Zhengyuan Liu, Alice Oh, and Jose Camacho-Collados. 2026. SemEval-2026 Task 7: Everyday Knowledge Across Diverse Languages and Cultures. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.

Anil Ramakrishna, Yixin Wan, Xiaomeng Jin, Kai-Wei Chang, Zhiqi Bu, Bhanukiran Vinzamuri, Volkan Cevher, Mingyi Hong, and Rahul Gupta. 2025. SemEval-2025 task 4: Unlearning sensitive content from large language models. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2584–2596, Vienna, Austria. Association for Computational Linguistics.

Marco Siino. 2024. All-mpnet at SemEval-2024 task 1: Application of mpnet for evaluating semantic textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 379–384, Mexico City, Mexico. Association for Computational Linguistics.

Ankit Vaidya, Aditya Gokhale, Arnav Desai, Ishaan Shukla, and Sheetal Sonawane. 2024. CLTeam1 at SemEval-2024 task 10: Large language model based ensemble for emotion detection in Hinglish. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 365–369, Mexico City, Mexico. Association for Computational Linguistics.

Raul Vazquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sanchez Vega, Alessandro Raganato, Jindřich Libovický, Jussi Karlgren, Shaoxiong Ji, Jindřich Helcl, Liane Guillou, Ona De Gibert, Jaione Bengoetxea, Joseph Attieh, and Marianna Apidianaki. 2025. SemEval-2025 task 3: Mu-SHROOM, the multilingual shared-task on hallucinations and related observable overgeneration mistakes. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2472–2497, Vienna, Austria. Association for Computational Linguistics.