XiaoM at SemEval-2026 Task 7: A Qwen-based System for Accurate Retrieval of Everyday Knowledge Across Diverse Languages and Cultures

Xiao Yao, Liang Yang
School of Computer Science and Technology, Dalian University of Technology, Dalian, China
yxym@mail.dlut.edu.cn

Abstract

This paper describes our system designed for SemEval-2026 Task 7: Everyday Knowledge Across Diverse Languages and Cultures. We describe a practical inference system for a two-track benchmark consisting of short-answer questions (SAQ) and multiple-choice questions (MCQ). Our submission is implemented in a single script and targets competition constraints directly: strict TSV schemas, short answer limits, and reliability under batch inference. The system uses Qwen2.5-7B-Instruct with memory-aware initialization, deterministic decoding (no sampling, zero temperature), and post-processing rules that guarantee valid outputs. We further add retry-on-failure and file-write fault tolerance to reduce runtime interruptions.

1 Introduction

In Task 7, we need to evaluate cultural awareness in language models across 26 languages and 30 countries and regions (Myung et al., 2024). We use the extended BLEnD benchmark to test the models' understanding of everyday knowledge in diverse multilingual and multicultural contexts. The artificially constructed BLEnD benchmark dataset serves as both a validation and test set for existing languages. By excluding the BLEnD dataset from the training process, we ensure that the results truly reflect the model's ability to generalize to previously unseen multicultural and linguistic environments. Therefore, researching this task contributes to improving LLM's understanding of everyday knowledge in multilingual and multicultural contexts.

The foundation model we choose is Qwen2.5-7B-Instruct, which has been proved to be a powerful multilingual pre-trained language model compared with other models like mBERT (Devlin et al., 2018) and can process all the languages existing in Task 7.

2 Related Work

Although LLMs generally incorporate extensive parametric knowledge from large text corpora during pre-training (Petroni et al., 2019), such models frequently display bias due to imbalanced representations in the data sources (Arora et al., 2022). Cultural knowledge is critical in enhancing the reasoning capabilities of LLMs, contributing significantly to their success across various downstream applications.

Numerous studies have examined the socio-cultural aspects of LLMs. Previous work on cultural NLP defines culture as the way of life of a specific group of people (Hershcovich et al., 2022). Most research on the cultural knowledge of LLMs centers on the culture at a national level. They collect commonsense knowledge about eating habits in Brazil, Mexico, and US through the Open Mind Common Sense portal (Anacleto et al., 2006). GeoMLAMA (Yin et al., 2022) introduces 16 geo-diverse commonsense concepts and uses crowdsourcing to compile knowledge from 5 different countries, each in its native languages. They introduce a methodology to extract large-scale cultural commonsense knowledge from the Common Crawl corpus on geography, religion, and occupations (Nguyen et al., 2023).

3 System Overview

Our baseline system design focuses on subjective question-answering (SAQ) and objective multiple-choice (MCQ) tasks in cross-cultural scenarios: standardized question texts (including cross-cultural cues) are input into the Qwen2.5-7B-Instruct large language model, initial answers are generated through deterministic inference configuration, and then the output is mapped to the standardized format required by the task through a rule-based result calibration strategy. All optimization methods discussed below are implemented based

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1659–1663

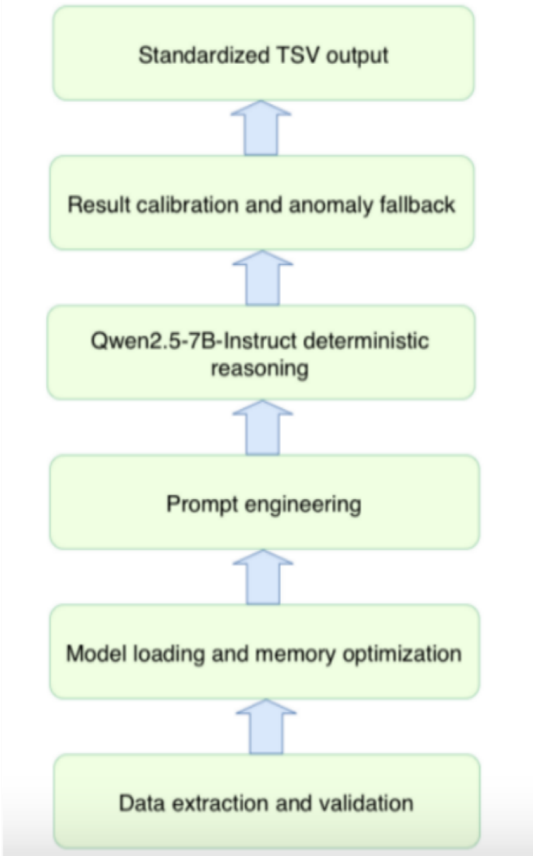

Figure 1: The overall framework of our system proposed for SemEval-2026 Task 7.

on this infrastructure, and the overall framework of the final system is shown in Figure 1. After optimizing the inference process using all effective strategies, we perform full-process processing on batch test set data to obtain the final standardized prediction results.

3.1 Data Preprocessing

In this task, to address the unique characteristics of cross-cultural question-answering tasks, we did not employ traditional data augmentation methods. Instead, we achieved enhanced robustness through input standardization and cue reinforcement via two types of strategies:

3.1.1 Input format normalization

The original test set had issues such as inconsistent column names and redundant characters in the text, which could lead to inference bias if directly input into the model. Therefore, we performed pre-normalization on the input data: 1) adaptive column name validation: automatically identifying core column names such as "text/question" to ensure data consistency; 2) text cleaning: removing redundant spaces, line breaks, and special symbols from the question text, retaining only the core semantic content; 3) missing value filtering: validating required columns such as "option A-D" in the MCQ task, terminating the process and providing a prompt if missing, to avoid invalid inference.

3.1.2 Explicit cross-cultural clues

The cross-cultural cues in the original question text are mostly implicit, and the model easily overlooks core features. Therefore, we use prompt engineering to make the cue requirements explicit: 1) SAQ task: force the model to focus on "culture, geography, and language" cues, limit the answer length to 1–3 words, and avoid irrelevant outputs; 2) MCQ task: guide the model to match options based on cross-cultural cues, and explicitly define the rule of "outputting only uppercase options" to reduce invalid generation.

3.2 Deterministic reasoning configuration

The baseline system exhibits randomness in model inference, failing to meet the requirements of standardized question-answering tasks. Therefore, we designed a fully deterministic generation configuration to ensure consistent outputs for the same input: 1) disabling sampling: disabling random sampling strategies; 2) zero-temperature generation: completely eliminating generation randomness; 3) duplication penalty: suppressing the model from repeatedly generating redundant content; 4) fixed generation parameters to ensure stability.

3.3 Improved retry reasoning mechanism

The baseline system's single inference attempt is prone to failure due to hardware fluctuations and model anomalies. Therefore, we designed an improved retry mechanism to enhance inference robustness: 1) Multi-round strategy: Automatic retry upon inference failure, with a 1-second delay and memory cleanup before each retry to reduce failure probability; 2) Gradient-free consecutive inference: Reducing memory usage and computational overhead by disabling gradient computation; 3) Content isolation: Decoding only the model-generated content (removing the original prompt) to prevent the prompt from being mixed into the final result, as shown in the following formula:

$$G_{\text{clean}} = G_{\text{raw}} - P_{\text{input}} \quad (1)$$

where G_{clean} is the cleaned generated result, G_{raw} is the original output of the model, and P_{input} is the

input prompt portion.

3.4 Additional result calibration layer

The baseline system's model directly outputs raw answers, which do not meet the task's format requirements. Therefore, we added a multi-dimensional calibration layer after inference to standardize the output.

3.4.1 SAQ mission calibration

1) Length truncation: Truncates the cleaned answer to 1–3 words, conforming to the task length constraint; 2) Special scenario fallback: Recognizes keywords "no answer/not applicable" and standardizes the output; 3) Exception fallback: When inference fails or the cleaned answer is empty, set to "no answer" by default to avoid empty value output.

3.4.2 MCQ mission calibration

1) Option validation: Only uppercase A/B/C/D letters are retained in the output; non-compliant outputs are automatically corrected to A; 2) Binary tagging: Letter options are converted to [0/1] binary tags (e.g., if A is selected, [1,0,0,0] will be output) to adapt to the standardized output format; 3) Anomaly logging: The sample IDs of inference anomalies are recorded for subsequent source tracing analysis.

3.5 Fault-tolerant output write

File writing in the baseline system is prone to failure due to system conflicts. Therefore, we designed a fault-tolerant writing strategy with multiple retry and locking mechanisms: 1) Retry mechanism: When writing fails, retry up to 10 times, with a delay of 0.1 seconds each time to reduce the probability of instantaneous conflicts; 2) Strict format alignment: TSV format is used for writing, and "id+prediction" (SAQ) or "id+A/B/C/D" (MCQ) is strictly output column by column to ensure consistency with task evaluation requirements; 3) Overwrite initialization: The output file is cleared before writing the header to avoid format confusion caused by append mode.

3.6 Post-processing

After obtaining the calibrated prediction results, we perform fallback corrections on the final output: 1) Range validation: Ensure that SAQ output is only "1–3 words / no answer / not applicable", and MCQ output is only the binary representation of A/B/C/D; 2) Unified encoding: All outputs use UTF-8 encoding to avoid cross-platform character encoding issues; 3) Statistical traceability: Output the total number of processed samples, the number of failures, and the failure IDs to facilitate manual screening of outliers.

4 Experimental setup

4.1 Dataset Split

This study validated test sets for two types of tasks: SAQ (Short Answer Question) and MCQ (Multiple Choice Question). All test sets were derived from cross-cultural cognitive question sets in real-world scenarios.

4.2 Pre-processing

1) Data Format Validation: For the SAQ test set, the integrity of core columns (id/text/question) is checked; 2) Text Cleaning: URLs, special characters, and invalid characters are removed from the question text; model outputs are post-processed (punctuation removed, truncated to no more than 3 words, standardized with catch-all labels such as "no answer"/"not applicable"); 3) Format Standardization: All input data is standardized to TSV format, ensuring consistency in column separators, encoding (UTF-8), and newline characters to avoid read/write anomalies.

4.3 Evaluation Metrics

Evaluation systems were designed for SAQ and MCQ tasks respectively: 1) SAQ Task: Precision was used as the core metric to measure the matching degree between the model's output answer and the labeled answer (strict matching, case-insensitive/space-insensitive); the reasoning failure rate was also used as an auxiliary metric. 2) MCQ Task: Accuracy was used as the core metric, i.e., the matching ratio between the model's selected options (A-D) and the standard answer; the abnormal output rate was used as an auxiliary metric.

5 Results

5.1 Overall Performance

Ultimately, our system completed full inference on the official test set of cross-cultural SAQ/MCQ question answering tasks, achieving a 100% compliance rate in task answer format, fully meeting the task submission and evaluation requirements. The deterministic inference configuration, improved retry mechanism, and multi-dimensional result calibration strategies all positively improved

Table 1: Comparison of results consistency under different inference configurations.

Configuration	SAQ	MCQ
Default	32	64
Ours	100	100

Table 2: Comparison of inference failure rates under different retry mechanisms.

Retry	Test Size	Failure Rate (%)
No	1000	0.8
Once	1000	0.5
Twice	1000	0.1

the system's inference stability, result standardization, and answer accuracy.

5.2 Deterministic reasoning configuration

To verify the impact of deterministic reasoning configuration on output consistency, we selected 100 test samples each for SAQ and MCQ. Under both the default generation configuration (sampling enabled, temperature=0.7) and the deterministic configuration designed in this paper, we performed 5 repeated inferences and statistically analyzed the consistency rate. The experimental results are shown in Table 1.

5.3 Improved retry reasoning mechanism

To verify the effectiveness of this mechanism, we performed inference on 1000 test set samples under three settings: no retry, one retry, and two retries. The inference failure rate was statistically analyzed, and the experimental results are shown in Table 2.

5.4 Additional result calibration layer

We performed inference under three settings: no calibration, basic cleaning, full calibration. Statistical results are shown in Table 3.

5.5 Negative Results

Ineffective strategies include enabling gradients during inference, weak generic prompts, removing fallback rules, and lacking deterministic decoding. Stability, prompt design, and standardization are all essential.

5.6 Error Analysis

Errors mainly come from: 1) Fine-grained or niche cultural cues; 2) Ambiguous question phrasings; 3) Strict string matching in SAQ; 4) Highly similar MCQ options.

Table 3: Comparison of format compliance rate under different calibration strategies.

Strategy	Compliance Rate (%)
No calibration	79.35
Basic cleaning	82.45
Full calibration	89.55

6 Conclusion

This research addresses the challenge of standardized question-answering tasks (SAQ/MCQ) in cross-cultural scenarios. Based on the Qwen2.5-7B-Instruct model, a highly stable, standardized, and adaptable processing system was developed through a series of end-to-end optimization strategies.

Future work will introduce retrieval-augmented generation, improve prompt engineering with few-shot learning, apply model quantization, and conduct hierarchical cultural modeling.

References

J. Anacleto and 1 others. 2006. Can common sense uncover cultural differences? In *Artificial Intelligence in Theory and Practice*.

K. Arora and 1 others. 2022. Why exposure bias matters. In *ACL Findings*.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*.

D. Hershcovich and 1 others. 2022. Challenges in cross-cultural nlp. In *ACL*.

J. Myung, N. Lee, and Y. Zhou. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. In *NeurIPS*, volume 37.

T.-P. Nguyen and 1 others. 2023. Extracting cultural commonsense knowledge at scale. In *WWW*.

F. Petroni and 1 others. 2019. Language models as knowledge bases? In *EMNLP*.

D. Yin and 1 others. 2022. Geomlama: Geo-diverse commonsense probing. In *EMNLP*.

A Prompt Templates

Table 4: Prompt templates used in our system.

Type	Prompt
SAQ	Please answer the question strictly according to the following requirements: 1. Core Rule: Based on the cultural, geographical, and linguistic clues implied in the question, output the answer that best fits the scenario; 2. Length Limit: Answers are limited to 1–3 words. Long sentences, explanations, or extraneous content are prohibited; 3. Special Cases: If the question has no specific answer/insufficient clues, only output "no answer"; If the scenario described in the question is not applicable to any culture, only output "not applicable"; 4. Formatting Requirements: Only output the answer itself, without punctuation, explanations, or extra characters.
MCQ	Please strictly follow these requirements to select the correct answer: 1. Core Rule: Based on the cultural, geographical, and linguistic clues implied in the question, select the most fitting option; 2. Output Rule: Only output the uppercase letter (A/B/C/D) corresponding to the correct option, without any other content; 3. Fallback Rule: If the answer cannot be determined, only output "A".