FLANS at SemEval-2026 Task 7: RAG with Open-Sourced Smaller LLMs for Everyday Knowledge Across Diverse Languages and Cultures

Liliia Bogdanova[1,†], **Shiran Sun**[2,†]
Natalia Amat Lefort[3], **Flor Miriam Plaza-del-Arco**[3], **Lifeng Han**[*3,4]
[1] Insilico Medicine AI Limited [2] University of Groningen
[3] LIACS, Leiden University [4] Leiden University Medical Center
[†] *co-first* alphabet order [*]LH: *Corresponding: l.han@liacs.leidenuniv.nl*

Abstract

This system paper describes our participation in the SemEval-2026 Task-7 "Everyday Knowledge Across Diverse Languages and Cultures". We attended two subtasks, i.e., Track 1: Short Answer Questions (SAQ), and Track 2: Multiple-Choice Questions (MCQ). The methods we used are retrieval augmented generation (RAGs) with open-sourced smaller LLMs (OS-sLLMs). To better adapt to this shared task, we created our own culturally aware knowledge base (CulKBs) by extracting Wikipedia content using keyword lists we prepared. We extracted both culturally-aware wiki-text and country-specific wiki-summary. In addition to the local CulKBs, we also have one system integrating live online search output via DuckDuckGo. Towards better privacy and sustainability, we aimed to deploy smaller LLMs (sLLMs) that are open-sourced on the Ollama platform. We share the prompts we developed using refinement techniques and report the learning curve of such prompts. The tested languages are English, Spanish, and Chinese for both tracks. Our resources and codes are shared via https://github.com/aaronlifenghan/FLANS-2026

1 Introduction

We present the system report for SemEval Shared Task 7 (Ousidhoum et al., 2026). Our contributions are: (1) We propose a sustainable and locally deployable RAG framework using open-sourced small language models for multilingual cultural question answering. (2) We construct a multilingual cultural knowledge base from Wikipedia and curated facts to support culturally grounded QA. (3) We perform a prompt ablation study demonstrating the importance of structured prompt design for multilingual factual QA. (4) We introduce a cascaded retrieval and model routing strategy that improves answer grounding while maintaining efficiency.

2 Related Work

I) RAG for Culture-Aware Generation LLMs have been known for their struggles when dealing with culturally specific content due to knowledge sparsity. To address this issue, there are related work using RAGs for culturally-aware text generation. For example, from **country** and **region** perspective, Lee et al. (2025) proposes a structured evaluation framework to assess how well LLMs represent and generate minority cultural knowledge — specifically focusing on Taiwanese Hakka culture. Other language and culture specific RAG systems including Islamic and Arabic (Alan et al., 2025; AbdelAziz et al., 2025; Faruk et al., 2025), Yoruba Culture and Language (African) (Joshua, 2024), etc. Similarly, focusing on specific domains such as **education**, LLMs' hallucination can perpetuate bias or outdated knowledge that are especially risky when generating learning content for vulnerable populations. Joseph et al. (2024) focused on refugee learners and explored RAGs as a strategy to improve the quality, contextual relevance, and cultural sensitivity of educational content.

In a **multilingual** setting, while RAG helps in knowledge-intensive tasks, it can also propagate and amplify biases present in retrieved documents, especially in cross-lingual settings where source document quality varies. Li et al. (2025) contributed a new dataset BordIRLines, containing territorial dispute descriptions paired with relevant retrieved *Wikipedia* documents (we also used for our own work in this shared task). It covers 49 languages, spanning many language families and resource levels, designed to test how well RAG systems handle culturally sensitive retrieval and generation across languages. To address **cost** of human annotated benchmark, Zhang et al. (2025) provided a scalable synthesis framework that combines hierarchical cultural knowledge with retrieval mechanisms for higher-quality, culturally grounded QA

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 869–878

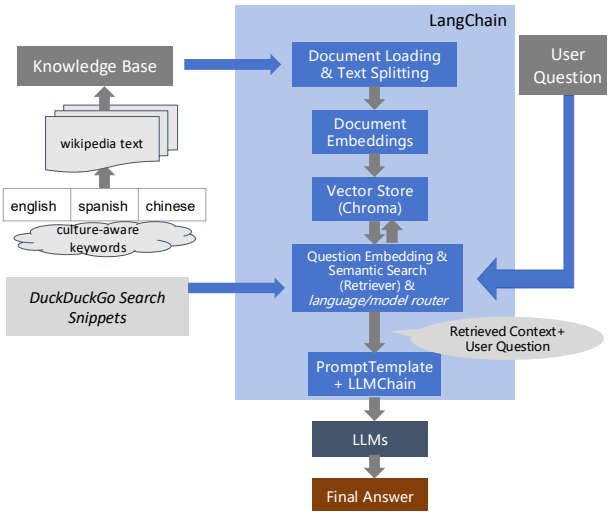

Figure 1: FLANS's RAG. *italic* indicating variations

generation. Unlike benchmarks that rely heavily on manual creation, CultureSynth demonstrates how taxonomy-guided RAG can produce large amounts of useful evaluation data across languages. Evaluation of 14 popular LLMs using CultureSynth shows larger models perform better on cultural QA tasks. A parameter threshold around 3B is necessary for basic cultural competence. Closely following this route, we aim to explore lower-cost, *sustainable*, *secure*, and smaller-sized LLMs on this task, for which we list some related literature below.

II) Sustainable and Secure AI Researchers have questioned the assumption that larger LLMs are always superior. Scaling-law studies by Kaplan et al. (2020) showed diminishing performance returns as the model size increases, while, instead the bigger the better, Han et al. (2024) demonstrated that smaller, better-trained models can match and even exceed over-sized counterparts, on the domain-specific machine translation (MT) task. NLP Researchers have started to highlight the substantial carbon and energy costs of large-score LM training and advocate efficiency-aware solutions (Strubell et al., 2019), e.g. model distillation, quantization, and lightweight open-source models (Jiao et al., 2020; Hu et al., 2023). Motivated by these findings, we explore open-sourced small language models (OS-sLLMs) and RAG to reduce computational cost, enable local deployment, and improve privacy on culturally grounded QA.

3 Methodology

The methodology design of our system is shown in Figure 1. The system is based on a multilingual knowledge base (KB) in which documents are split into small text chunks, encoded using a multilingual embedding model, and indexed in a local Chroma vector store, as shown in Figure 1 (upper-half). At inference time, the user question is embedded in the same vector space and matched against the index to retrieve the top-k most relevant passages via semantic search. The retrieved context is then combined with the original question and inserted into a structured prompt template, which is passed through the LLMChain to generate the final answer. The SLMs are instructed to output only a short factual response, ensuring concise and consistent outputs. short factual answer. This design enables the model to access culture-specific information while keeping the overall system lightweighted and compatible with local hardware.

The design of our system is **motivated** by the specific characteristics of the BLEnD shared task, which evaluates multilingual and culture-specific factual knowledge across diverse languages. Here we adopt a small-language-model (SLM) type. Small models alone lack sufficient parametric knowledge for culturally grounded questions, but their behaviors are expected to become more stable and predictable when combined with external evidence. Therefore, our system integrates: (1) lightweight multilingual SLMs suitable for local deployment, (2) structured prompt design to enforce short, same-language answers, and (3) a (RAG) module using a curated external knowledge base. This architecture aims to balance accuracy, multilingual robustness, and computational feasibility while providing a transparent and extensible/adaptable framework for evaluating cultural question answering. As a variation of our system (RAG-WEB), we added online search component via DuckDuckGo and language/model router function in *italic*.

3.1 Language and Model Selection

We first define the language scope of our system, which directly informs our model selection under local deployment constraints. Our development process focused on the three languages most relevant to our team and to the early stages of the BLEnD dataset: Chinese, English, Spanish. These languages were used to test model behavior, prompt stability, and retrieval quality in a multilingual setting.

To identify a suitable model for the BLEnD short-answer questions, especially for Chinese

queries, we first compared four small language models available in Ollama that match our local hardware constraints: Phi-4-Mini, Llama 3.2 (3B), DeepSeek-R1 (7B), and Gemma 3 (4B). We asked each model to generate answers for a subset of BLEnD SAQ items and evaluated them based on accuracy, response stability, and runtime efficiency. Among these candidates, Gemma 3 showed the most balanced performance, providing higher accuracy on Chinese factual questions and noticeably faster, more consistent output. Based on these observations, we selected Gemma 3 as the primary model for our RAG-BASE system.

In the *variation* version (RAG-WEB), we explored two models `mistral:7b` and `deepseek-llm:67b`. The models are chosen dynamically using the language identifier embedded in each of the BLEnD questions' ID. For the majority of inputs that are not Chinese, the `mistral:7b` model is used. This is because it strikes a good balance between multilinguality, inference speed, and local deployability. However, experiments have revealed that it is not very effective for Chinese queries in region-specific contexts. For this problem, a *deterministic routing* approach is adopted. If the language ID is one of the Chinese variants (such as `zh-CN`, `zh-TW`, `zh-SG`), the system switches to using the `deepseek-llm:67b` model. This is because it is much larger but performs better for Chinese factual reasoning and named entity recognition.

3.2 Prompt Development

We evaluate three prompt variants of increasing structural and cognitive constraint, denoted as Minimal Prompt (**MP**) and Refined Prompts (**RP-v1** and **RP-v2**). MP served as a baseline and used a minimal instruction, asking the model to produce one short, correct answer in the same language as the question, without additional guidance on format or reasoning. For Refined Prompts (RPs), we explore different kinds of techniques such as persona, format instruction, Chain-of-Thought (CoT) and perspective-aware prompting (Ren et al., 2025; Romero et al., 2025). Main design elements of RP-v1 include:

- Assigning the model the persona of being a factual multilingual assistant for a question-answering benchmark.
- Enforcing a strict output format instruction that produces only one concise answer in the same language as the question, without any explanations or extra text.

The goal of RP-v1 is to make the model's output more concise and consistent, and to reduce common errors such as unnecessary words or language switching, which small language models sometimes produce in multilingual settings. RP-v2 introduces a perspective-aware step inspired by PA-ISP, incorporating implicit CoTs reasoning to guide the model's internal analysis before producing the final answer (Ren et al., 2025). In this version, main design elements include, in addition to the persona and strict output format from RP-v1: a) Introducing a structured self-guided reflection phase before answering, b) Guiding the model to think internally in several structured steps, including question analysis, information focus, answer strategy, and error avoidance.

Among the Refined Prompts (RP-v1 and RP-v2), the difference is that RP-v2 adds an explicit internal analysis stage before answer generation. It encourages the model to examine the question from multiple perspectives, such as language identification and information type, before deciding on the final answer. This design aims to improve answer accuracy and consistency in multilingual settings, especially for short-answer questions where the model must select a single precise fact.

3.3 Knowledge Base Constructions

For **RAG-BASE** Knowledge Base Construction (**KBC**), we build a multilingual knowledge base in Chinese, English, and Spanish to support the question-answering task. The KB combines two sources: (1) Wikipedia content and (2) manually curated cultural facts. Wikipedia articles are selected to cover common types of general-knowledge questions, including national symbols, geography, history, culture, and society. For each topic, we extract sentences from introductory summaries, which typically contain clear and concise factual information. These segments, either sentences or short paragraphs, are used as individual KB entries, resulting in approximately 700 entries from Wikipedia. To ensure reliable coverage of important and frequently asked cultural facts, we additionally include around 200 manually curated statements, such as official currencies or well-known national associations. All KB entries are embedded using the OllamaEmbeddings model and indexed with

Chroma for vector-based retrieval. [1] For RAG-BASE system, if the model retrieve some knowledge, it will print the sentence out, otherwise it prints “NULL".

3.4 Creating Pseudo Ground Truth

Since the official BLEnD gold answers cannot be used for training or tuning, we created a small pseudo ground truth set to support model development and system debugging. This set was constructed using two complementary methods. First, a small subset of Chinese, English and Spanish questions was manually annotated by native speakers or near-native speakers to ensure correctness. Second, we used GPT-4 to generate initial candidate answers, which were then manually checked and corrected via post-editing. This hybrid approach allowed us to build a reliable reference set without introducing data leakage from the official test labels. The pseudo ground truth was used only for internal evaluation of prompt quality, model selection, and RAG behavior, providing a controlled way to assess system changes during development.

3.5 Language Routing

Given BLEnD’s wide range of languages and cultural variants, we use a deterministic language routing approach. The language code embedded in each question ID (e.g., `es-MX`, `en-SG`, `zh-CN`) determines both the country-level knowledge base and the LLM used for answer generation. This means that Spanish queries about Mexico pull Mexican cultural information, while English queries about Singapore pull Singapore-specific sources. For Chinese variants, this routing also activates the `DeepSeek` model, which tends to give more reliable results than the lighter alternatives.

4 System Development and Validation

4.1 Experimental and Evaluation Setup

For RAG-BASE: we use Gemma-4B and run all experiments locally on a CPU-based laptop (Intel i5-10210U, 16 GB RAM). No GPU or external computing infrastructure is involved. This highlights the feasibility of lightweight and sustainable deployment.

For RAG-WEB: we run experiments for MCQ questions locally using Ollama with Mistral and DeepSeek models on a MacBook Pro (Apple M1, 16 GB RAM, 512 GB SSD). All computations were performed on CPU-based Apple Silicon without discrete GPU or external infrastructure. This demonstrates that competitive multilingual cultural QA performance can be achieved using modest consumer hardware, supporting sustainable and accessible AI deployment.

We use the standard BLEnD **evaluation protocol** for system evaluation. For the Short Answer Question (SAQ) track, a system prediction is said to be correct if it matches any of the human reference answers for a given question. For the Multiple Choice Question (MCQ) track, correctness is measured by matching a system prediction with a ground truth label. The final score is reported as accuracy averaged across all language tracks. More details about the evaluation protocol can be found in (Myung et al., 2024).

4.2 Ablation Study of Prompts

Since our system only covers three languages (English, Spanish, and Chinese), the remaining languages contribute zero scores using the official evaluation platform, resulting in a lower overall score than the actual performance on the evaluated languages. To obtain a clearer view, we report results for the selected languages only. For each language, scores are averaged over regional variants (e.g., `zh-CN` and `zh-SG` for Chinese). Detailed per-language results for both Track 1 (SAQ) and Track 2 (MCQ) under different prompt versions are shown in Table 1 using our RAG-BASE system. Figure 2 further visualizes the three-language average scores across prompt versions.

For Track 1, performance consistently improved from MP to RP-v2, suggesting that both explicit formatting instruction and perspective-aware prompting are beneficial for open-ended short-answer generation, where the answer space is relatively unconstrained. In contrast, Track 2 performance peaked at RP-v1 and slightly declined with RP-v2. This suggests that for multiple-choice questions, where the answer space is already tightly constrained by predefined options, additional internal reflection may introduce unnecessary cognitive overhead and interfere with option-level matching.

Overall, the prompt ablation suggests that most performance gains come from explicit external constraints on output format and multilingual behavior. While self-guided reflection can further improve performance on open-ended short-answer ques-

[1] we share our KB of wiki extraction and manually written facts at our page https://github.com/aaronlifenghan/FLANS-2026

Language	Track 1 (SAQ)			Track 2 (MCQ)		
	MP	RP-v1	RP-v2	MP	RP-v1	RP-v2
English (en)	17.14	24.29	37.14	82.86	82.86	82.86
Spanish (es)	4.17	35.00	47.50	80.83	85.00	70.00
Chinese (zh)	27.14	41.43	48.57	65.71	82.86	68.57
Average	16.15	33.57	44.40	76.46	83.57	73.81

Table 1: RAG-BASE Prompt ablation results by language and task.

Language	Track 1 (SAQ)			Track 2 (MCQ)	
	RP-v1	RP-v2	RP-v1 (no local DB)	RP-v1	RP-v1 (no local DB)
English (en)	16.67	0.00	16.67	83.33	83.33
Spanish (es)	33.33	0.00	33.33	61.11	72.22
Chinese (zh)	33.33	8.33	66.67	91.67	83.33
Average	27.78	2.78	38.89	78.70	79.63

Table 2: RAG-WEB Prompt ablation results by langauge and task.

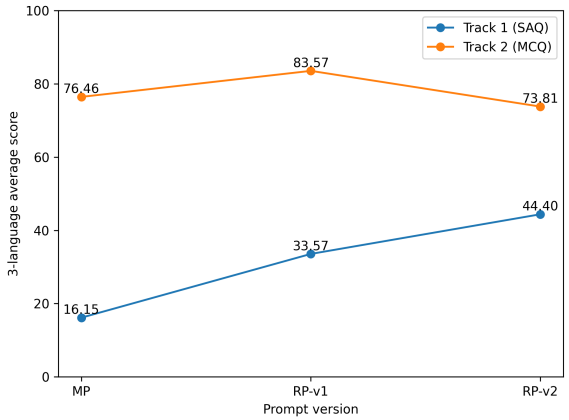

Figure 2: RAG-BASE Learning curves of three prompt ablation averaged over three-languages (en, es, zh).

tions, its benefits are limited for tasks with a tightly constrained answer space, such as multiple-choice questions. Based on these observations, we adopt RP-v1 as the default prompt in our RAG-BASE system, as it provides a strong balance between performance, simplicity, and robustness across both tracks.

For **RAG-WEB**, as shown in Table 2, scores are weighted averages over regional variants based on the number of evaluation questions. The evaluation on Chinese has much improved scores for Track 1 (SAQ) using RP-v1 achieving 66.67 vs the highest score 48.57 from RAG-BASE, although the other two languages have dropped scores. Here, "no local DB" means only using online search duckduckgo and wikipedia. For Track2, using local database, Chinese further improved the score from 83.33 to 91.67, but Spanish score decreased 10 absolute points (72.22 to 61.11), which means that the Spanish DB might have introduced noise. This leads to future work to explore DB quality control for RAG systems. Overall, we observe that performance on MCQ is consistently higher than on SAQ across all languages, reflecting the relative difficulty of free-form short-answer generation. Chinese achieves the strongest MCQ performance, benefiting from language-specific model routing to a larger LLM, while Spanish shows comparatively stronger gains on SAQ after aggregation.

4.3 Submission to SemEval-Test

For official submission to SemEval Task 7, we submitted three systems using RAG-BASE, RAG-WEB, and RAG-mix, where RAG-mix used RAG-BASE for SAQ and RAG-WEB for MCQ, all using the prompt RP-v1.

5 Conclusions and Future Work

In this system report, we described our FLANS system architecture design and model development for SemEval Shared Task 7 "Everyday Knowledge Across Diverse Languages and Cultures". For sustainable and secure AI development, we explored retrieval-augmented generation (RAG) using open-sourced smaller LLMs and Wikipedia extracted local knowledge base, together with online live search engine using language and model routing. We shared our detailed prompts developed and our codes for open-sourced research purposes. Future work includes improving knowledge base quality control, expanding language coverage, and further optimizing small model performance.

6 Limitations

There are a few limitations from our current work: 1) The coverage gap is apparent where we only focused 3 languages (EN/ES/ZH) out of the full BLEND set. This can make the official evaluation scores difficult to interpret and not comparable to other systems by default. 2) RAG-WEB results need more explainability, especially when it gives 0 scores on English and Spanish SAQ task for our prompt ablation studies. Further evaluations can include interpreting when and how RAG-WEB helps vs hurts. 3) Methodologically, the two RAG systems used different models; however, we plan to unify the comparisons in the future across two models, as well as adding some comparisons with strong baselines e.g. GPT-4 and other SemEval systems. 4) Other in-depth analysis and exploration would be needed such as: cascaded retrieval + routing, KB coverage extension, ablation analysis on retrieval quality, KB design, model choice, routing decision, and grounding evaluation.

Disclaimer The opinions and conclusions expressed in this paper are those of the authors (LB) and do not necessarily reflect the views of Insilico Medicine AI Limited.

Acknowledgments

We thank the reviewers for valuable comments on our work.

References

Abdelaziz Amr AbdelAziz, Mohamed Ahmed Youssef, Mamdouh Mohamed Koritam, Marwa Eldeeb, and Ensaf Hussein. 2025. Arabic mental health question answering: A multi-task approach with advanced retrieval-augmented generation. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 192–197.

Ahmet Yusuf Alan, Enis Karaarslan, and Ömer Aydın. 2025. Improving llm reliability with rag in religious question-answering: Mufassirqas. *Turkish Journal of Engineering*, 9(3):544–559.

K.M.Tahlil Mahfuz Faruk, Mushfiqur Rahman Talha, H. M. Kawsar Ahamad, Mohammad Galib Shams, Nabil Mosharraf Hossain, Syed Rifat Raiyan, Md Kamrul Hasan, Hasan Mahmud, and Riasat Islam. 2025. ADAB: A culturally-aligned automated response generation framework for islamic app reviews by integrating ABSA and hybrid RAG. In *5th Muslims in ML Workshop co-located with NeurIPS 2025*.

Lifeng Han, Serge Gladkoff, Gleb Erofeev, Irina Sorokina, Betty Galiano, and Goran Nenadic. 2024. Neural machine translation of clinical text: an empirical investigation into multilingual pre-trained language models and transfer-learning. *Frontiers in Digital Health*, 6:1211564.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2023. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Tibakanya Joseph, Nakayiza Hellen, and Ggaliwango Marvin. 2024. Retrieval-augmented llms for culturally sensitive learning content in refugee education. In *Congress on Intelligent Systems*, pages 425–442. Springer.

Adejumobi Monjolaoluwa Joshua. 2024. Improving question-answering capabilities in large language models using retrieval augmented generation (RAG): A case study on yoruba culture and language. In *5th Workshop on African Natural Language Processing*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Hung-Shin Lee, Chen-Chi Chang, Ching-Yuan Chen, and Yun-Hsiang Hsu. 2025. Evaluating cultural knowledge processing in large language models: a cognitive benchmarking framework integrating retrieval-augmented generation. *The Electronic Library*, pages 1–22.

Bryan Li, Fiona Luo, Samar Haider, Adwait Agashe, Siyu Li, Runqi Liu, Miranda Muqing Miao, Shriya Ramakrishnan, Yuan Yuan, and Chris Callison-Burch. 2025. Multilingual retrieval augmented generation for culturally-sensitive tasks: A benchmark for cross-lingual robustness. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4215–4241, 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, and 1 others. 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, and 11 others. 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.

Lei Ren, Y. M. Ng, and L. Han. 2025. Malei at MultiClinSUM: Summarisation of Clinical Documents using Perspective-Aware Iterative Self-Prompting with LLMs. *MultiClinSUM Shared Task at the 13th BioASQ Workshop with the CLEF conference*.

Pablo Romero, Libo Ren, Lifeng Han, and Goran Nenadic. 2025. The manchester bees at peranssumm 2025: Iterative self-prompting with claude and o1 for perspective-aware healthcare answer summarisation. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 340–348.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3645–3650.

Xinyu Zhang, Pei Zhang, Shuang Luo, Jialong Tang, Yu Wan, Baosong Yang, and Fei Huang. 2025. CultureSynth: A hierarchical taxonomy-guided and retrieval-augmented framework for cultural question-answer synthesis. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10448–10467, Suzhou, China. Association for Computational Linguistics.

A RAG-BASE and RAG-WEB

The two RAG systems we developed before merging are presented in Figure 3 and 4.

The steps to run RAG-BASE:

- 1st: run prompt; there is no output but just to set it up.
- 2nd: run RAG; it might lead to some output.
- 3rd: print answer from LLM-w-RAG.

A.1 RAG-WEB Design Motivation

The BLEnD shared task tests culturally grounded and country-specific knowledge in a wide variety of languages and geographic locations. A large proportion of the questions involve common practices, local customs, institutions, and/or facts that are particular to a geographic area and are not expected to be encoded in the parametric memory of a small language model. On the other hand, the BLEnD shared task does not allow task fine-tuning, which calls for an inference strategy that boosts grounding while maintaining low computational costs and reproducibility.

In light of the above-mentioned issues with the BLEnD shared task, we have devised a cascaded retrieval-augmented inference strategy that is optimized for local execution. Instead of depending on a single source of information, the model combines different channels of evidence with a fallback strategy: **web search** with low overhead, local knowledge base with country-specific information from Wikipedia (optional), and responses from a parametric model when applicable. The model prioritizes precision and format correctness over recall and explicitly allows abstention through a <NO_ANSWER> output when evidence cannot be established. The final model strikes a balance between local search and quality of answers due to a trade-off observed during development: local search improves recall but hurts the quality of answers for culturally nuanced short answers.

A.2 RAG-WEB KB

For each country included in BLEnD, a brief set of highly informative Wikipedia pages regarding country overviews, culture, cuisine, tourism, history, etc., was automatically compiled. This collection was then split into overlapping text segments using a character-based text splitter, which ensures semantic coherence. Each text segment was then embedded by `OllamaEmbeddings` (Mistral-based) before being stored in a country-specific Chroma vector database, `db_{country}`.

This structure enables efficient semantic retrieval, ensuring that culturally relevant data remains geographically relevant. Notably, the local knowledge bases are used as optional evidence sources, not as a source of ground truth, with access restricted at runtime.

A.3 RAG-WEB vs Traditional RAG

While the methodology takes its cue from conventional RAG frameworks, the proposed system differs in that it does not strictly follow a one-pass retrieval and generation strategy, as in the case of conventional RAGs, and instead adopts a cascaded inference strategy that is more suitable for the BLEnD task. Each country in the dataset is associated with an optional local knowledge base, implemented as a separate Chroma vector database. The databases are built on a curated subset of relevant Wikipedia pages, including those related to national overview, culture, cuisine, tourism, and history.

In terms of its inference strategy, the proposed system follows a multi-stage decision process. First, the question ID is tokenized and analyzed to determine the language and country code, which in turn determines the language model and, if enabled, the country-specific local knowledge base. The system next determines if a direct model-only response is appropriate, a strategy primarily reserved for encyclopedic types of queries and not culturally informed short-answer queries, in which case an unguided generation process is not reliable.

In terms of retrieval, if evidence is requisite, the system uses a prioritized cascaded strategy. First, web search using `DuckDuckGo` is attempted, as this search engine is more likely to return concrete and contextually relevant information snippets relevant to everyday cultural queries (we consider the top-8 candidates). If this fails, the system will attempt a semantic similarity search using the country-specific webpage from Wikipedia summary (the free text in the top the of the page).

At each stage, retrieved evidence is injected into the prompt and the model is asked to answer only if the answer is explicitly supported by the context. If no stage produces a valid answer, the system outputs `<NO_ANSWER>`.

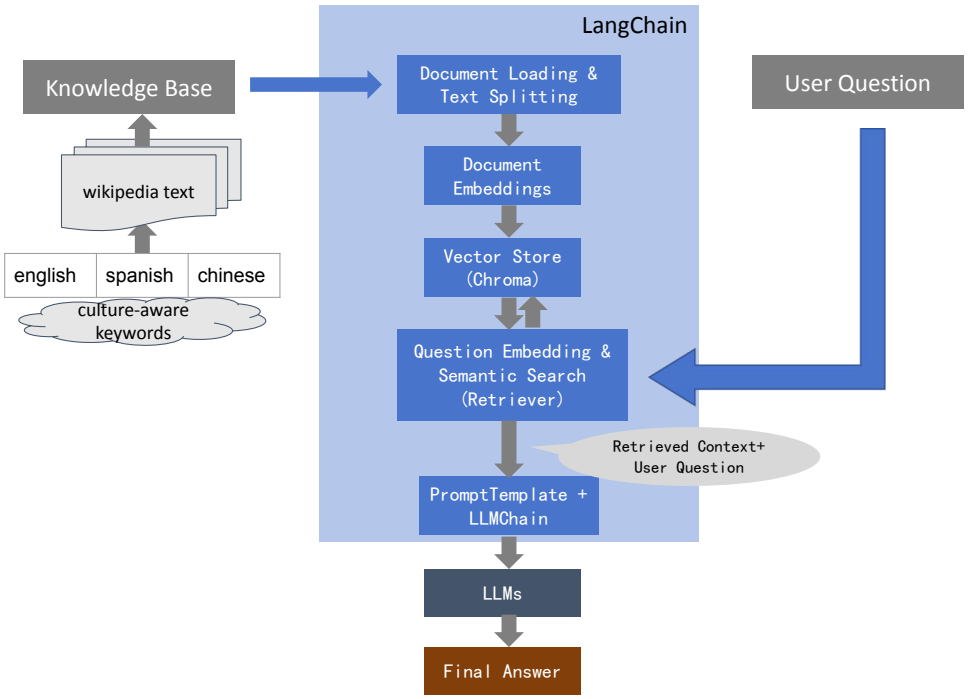

Figure 3: RAG-BASE pipeline using smaller LLMs favoring Gemma3.4b - keywords based KE - then land to KB

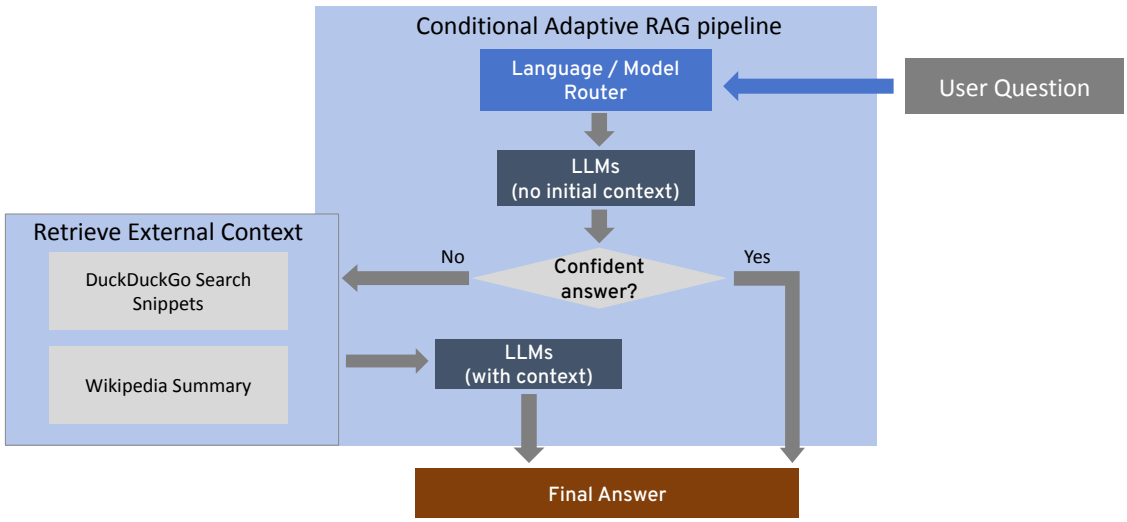

Figure 4: RAG-WEB pipeline using mistral:7b and deepseek-llm:67b (bigger) - Conditional adaptive RAG

A.4 RAG-WEB inference

Our final stage of inference utilizes a combination of retrieval from multiple external sources with dynamic routing of the language model. Depending on the type of query and availability, the system may retrieve supporting evidence from live web search, country-specific local knowledge bases, or Wikipedia summaries.

In the presence of local knowledge bases, queries are embedded in the same vector space as country-specific `Chroma` databases to retrieve semantically relevant cultural documents. Local knowledge retrieval, however, is not a necessary component, as empirical analysis showed that local retrieval can add noise to culturally grounded short-answer questions.

For most queries, web search through `DuckDuckGo` is preferred, especially when queries are about general practices, cuisine, or institutions. Wikipedia summaries are used as a fallback option when queries are encyclopedic or historical in nature. Retrieved passages are embedded into a prompt template, with the model being asked to respond only when the answer is explicitly supported by the evidence provided.

B Detailed Prompts

This appendix presents the three prompt variants used in our prompt ablation study.

B.1 Minimal Prompt (MP)

```
"""
Our goal is to give one short, correct answer for
↪  each question in its original language.

"""
```

B.2 Refined Prompt-v1 (RP-v1): Instructions + Persona

```
"""
You are a factual multilingual assistant for a
↪  question-answering benchmark.
Your goal is to give one short, correct answer
↪  for each question in its original language.

Instructions:
- Read the question carefully.
- Respond ONLY with the concise answer — a word,
↪  number, name, or short phrase.
- Do not include explanations, reasoning, labels,
↪  or extra words.
- If the question asks for a person, place, or
↪  date, give only that entity.
- Keep the answer in the SAME language as the
↪  question
  (Chinese → Chinese, English → English, Spanish
  ↪  → Spanish).

Now answer the following question.

"""
```

B.3 Refined Prompt-v2 (RP-v2): Persona + Perspective-aware + CoTs

```
"""
You are a factual, multilingual assistant for a
↪  question-answering benchmark.
Your goal is to produce ONE short, correct answer
↪  for each question in its original language.
Before answering, please think about the task:

1. Question Analysis:
- Identify the language of the question.
- Identify what type of information is being
↪  asked (person, place, date, object, concept,
↪  number, or other).

2. Information Focus:
- Determine the single factual element required
↪  to answer the question.
- Ignore any irrelevant or descriptive details.

3. Answer Strategy:
- Recall general world knowledge relevant to the
↪  question.
- Prefer the most standard, widely accepted
↪  answer.
- Avoid over-specific or explanatory phrasing.

4. Error Avoidance:
- Do NOT include explanations, reasoning, or
↪  extra words.
- Do NOT translate or restate the question.
- Do NOT include multiple candidates or
↪  alternatives.

After this internal analysis, provide ONLY the
↪  final answer.

Answering Rules:
- Output a single short answer (a word, name,
↪  number, or short phrase).
- Keep the answer in the SAME language as the
↪  question.
- Do not include labels, punctuation, or
↪  additional text.
- If uncertain, give your best plausible short
↪  answer based on general knowledge.

Now answer the following question.

"""
```