Models Without Borders at SemEval-2026 Task 7: Bridging Cultural Contexts with Search-Grounded QA

Swetha Krishna Sriram
BITS Pilani, Hyderabad Campus
kswetha036@gmail.com

Nirupama Sekar
Independent Researcher
sekarnirupama@gmail.com

Abstract

We present our submission to SemEval-2026 Task 7, focusing on the MCQ track, where models must identify culturally specific answers across language-region locales. Our system augments a compact open-source model with locale-targeted web retrieval at inference time, requiring no task-specific fine-tuning, and places 10th on the leaderboard. Beyond the submitted system, we explore how retrieval depth and search localization affect performance across locales, finding that localizing search parameters meaningfully shifts the geographic composition of retrieved sources and that gains from retrieval are most pronounced for lower-resource locales. We also investigate whether culturally informed prompt framing can complement retrieval, finding that it does, but only when grounding context is present. Taken together, our results point to inference-time web grounding as a practical path toward more culturally aware NLP under resource constraints.

1 Introduction

Large Language Models are deployed today in applications across diverse cultural contexts and used by users around the world, making it imperative to investigate their cultural knowledge and mitigate biases. Models trained predominantly on English and high-resource web content tend to reflect Western-centric perspectives, performing well on cultures that are heavily represented online while struggling with those that are not. Addressing this gap requires evaluation on benchmarks that probe culture-specific everyday knowledge.

SemEval-2026 Task 7 (Ousidhoum et al., 2026; Ghosh et al., 2026) provides such a benchmark by presenting an extended version of the BLEnD dataset (Myung et al., 2024), evaluating cultural awareness across locales. We participate in the multiple-choice question (MCQ) track, in which models must identify the culturally appropriate option for a given question, with all questions posed in English.

Our team, **Models Without Borders**, presents a retrieval-augmented system that grounds predictions with results from locale-targeted web search at inference time. Building on small open-source models without any task-specific fine-tuning, we use geographic and language parameters derived from locale identifiers to bias search results toward regionally relevant sources, and explore culturally informed prompt framing as a complementary intervention. Our system places 10th overall on the MCQ leaderboard with an accuracy of 78.81%. Beyond the submitted system, our experiments shed light on how retrieval depth, model choice, and prompt framing interact with cultural knowledge coverage, and what per-locale performance reveals about the uneven distribution of cultural knowledge on the web.

2 Background

2.1 Task Description

BLEnD (Myung et al., 2024) is a hand-crafted benchmark used popularly in literature to evaluate LLMs' cultural knowledge across 16 countries and 13 languages, including low-resource languages such as Amharic, Assamese, and Hausa. BLEnD focuses on everyday knowledge about daily practices and lifestyles, such as foods and customs, that is traditionally underrepresented in standard pretraining data.

SemEval-2026 Task 7 extends BLEnD to over thirty language-region locales as a shared task. The MCQ track, which our team has participated in, presents English-language questions where the model must identify the culturally correct answer for a target locale from among distractors often drawn from other cultures. A sample question from the MCQ track is given below.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3250–3257
July 3-4, 2026 ©2026 Association for Computational Linguistics

ID: `ga-IE_0001`
Question: *At what age do kids start nursery in Ireland?*
A) 4 **B)** 5 **C)** 6 **D)** 7

Cultural knowledge is by nature distributed unevenly across the training data used by LLMs - some regions have rich, accessible documentation of their everyday practices, while others do not. Rather than relying solely on what a model has internalized during pretraining, we employ a Retrieval-Augmented Generation (RAG) approach, augmenting our prompts with locale-aware web search results retrieved at inference time.

2.2 RAG for Cultural QA

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) integrates non-parametric documents into the generation pipeline, improving performance on knowledge-intensive tasks by conditioning predictions on retrieved context. This is a natural fit for culturally grounded QA, where parametric knowledge is unevenly distributed across locales by construction.

Subsequent work explores live web grounding, where search results are prepended to prompts at inference time, demonstrating gains on culturally grounded QA (Lertvittayakumjorn et al., 2025). Parallel efforts replace open web search with structured grounding. Nyandwi et al. (Nyandwi et al., 2025) construct culturally aligned supervision by retrieving and synthesizing multilingual knowledge from Wikidata, enabling more controlled coverage across languages. Zhang et al. (Zhang et al., 2025) propose a taxonomy-guided retrieval and synthesis framework that organizes cultural concepts hierarchically before generation, highlighting the role of retrieval quality and knowledge structuring in downstream performance.

Collectively, this body of work suggests that external grounding can mitigate uneven cultural coverage, but existing approaches often rely on curated resources or additional training procedures. In contrast, we examine lightweight inference-time interventions —localized web retrieval and culturally oriented prompt framing — to understand how external evidence interacts with parametric knowledge in compact open-source models under realistic compute constraints.

Motivated by this landscape, we frame our experiments around the following research questions:

- **RQ1:** What is the impact of locale-aware web search grounding on compact open-source models?
- **RQ2:** How does retrieval depth (Top k web search results) affect performance across diverse language-region locales?
- **RQ3:** Do locale-specific search parameters meaningfully shift the geographic provenance of retrieved sources?
- **RQ4:** Do culturally oriented prompt framing strategies complement retrieval in improving performance on culturally grounded QA?
- **RQ5:** Can geographically and linguistically conditioned web retrieval serve as a lightweight alternative to training-intensive cultural adaptation?

3 Methodology

Our system follows a RAG based approach, where we use live web search as the retrieval engine over a static document corpus. For each unique query in the input set, we retrieve locale-targeted web search results and use them to construct a grounded prompt. This is then fed to a language model for answer prediction. Our approach is described in detail below.

3.1 Web Retrieval

Each instance in the dataset carries an identifier of the form `{language}-{REGION}_{index}` (e.g., `ga-IE_0056`). We parse this to extract a language code (`hl`) and a country code (`gl`), which are passed as parameters to the Serper.dev API endpoint alongside the question text as the search query. Using `gl` and `hl` biases Google's results toward the regional and linguistic context of the question rather than returning a generic result set. The question text is used as the query without modification. This keeps the pipeline simple but may introduce noise in MCQ settings due to the presence of distractor options; addressing this through query reformulation is left to future work.

The API returns a JSON response containing an array of result objects. From each result we extract the `title` and `snippet` fields, which are formatted into a numbered list and prepended to the model prompt as grounding context (Appendix A). We retrieve and cache the top 10 results per query, from which we draw the top *k* results for each experimental condition. During this process, we observed that some responses included snippets originating from the original BLEnD dataset page on Hugging Face (nayeon212/BLEnD). To comply with competition rules, whenever such a result appeared within the top *k* positions, we excluded it and substituted the *(k+1)*-th result in its place.

3.2 Prompt Construction

The prompt consists of four elements in order: (1) a task framing statement instructing the model to select the answer option specific to the region or culture mentioned in the question; (2) an explicit note that the distractor options are drawn from *other* regions and cultures; (3) the formatted search context block; and (4) the question followed by the four labelled answer options. The model is constrained to respond with only a single letter from {A, B, C, D}. The full prompt template is reproduced in Appendix B.

3.3 Inference

We use Qwen2.5-7B-Instruct (Yang et al., 2024) as our primary backbone for the submitted system. For ablation experiments we additionally evaluate Aya-Expanse-8B (Dang et al., 2024), a multilingual model of comparable size. All inference is run with vLLM (Kwon et al., 2023), which enables high-throughput generation via continuous batching and CUDA graph optimization. Experiments are conducted on a single NVIDIA A100 80GB GPU. All decoding is greedy with a maximum of 5 new tokens.

4 Results and Discussion

4.1 Submitted System

Our submitted system - Qwen2.5-7B-Instruct with Top $k = 5$ search results given as context in prompt - achieves an overall accuracy of 78.81%, placing 10th on the MCQ leaderboard. Per-locale results are reported in Table 1. Performance varies substantially across locales, ranging from 62.47% on `az-AZ` (Azerbaijan) to 92.94% on `es-EC` (Ecuador). Broadly, locales associated with higher-resource languages tend to score higher, while locales such as `az-AZ`, `ha-NG`, `ko-KP`, and `su-JB` cluster at the lower end.

4.2 Effect of Search Localization on TLD Composition

To assess whether geographic and language parameters meaningfully alter the composition of retrieved search results, we compare two retrieval configurations: a *general* configuration using US-based settings, and a *localized* configuration specifying country-appropriate `gl` (geolocation) and `hl` (host language) parameters. We measure the proportion of local top-level domains (TLDs) in the top-k results as a proxy for geographic relevance of retrieved sources. We used tldextract to normalize retrieved URLs and extract registered domains based on the Mozilla Public Suffix List. This enabled robust aggregation at the root-domain level and avoided misclassification of multi-level country-code TLDs (e.g., co.uk, ac.in).

The difference is substantial. Under general search settings, an average of 6.86% of retrieved URLs per query carry a local TLD. Under localized settings, this rises to 18.47%, a mean gain of 11.61 percentage points. For some queries, localization increases local domain representation by up to 90%(Appendix C).

These findings have a direct implication for culturally grounded retrieval, demonstrating that generic US-centric search is a poor proxy for local information access. For locales where cultural knowledge is primarily encoded in region-specific sources, failing to specify `gl` and `hl` parameters risks surfacing globally prominent but geographically mismatched documents. The *Localization Gain* metric - the per-query increase in local TLD proportion - further serves as a natural signal of locale-specificity. We see that questions with high gain are those for which local context is structurally necessary, while near-zero gain questions likely concern globally generic facts amenable to any search configuration.

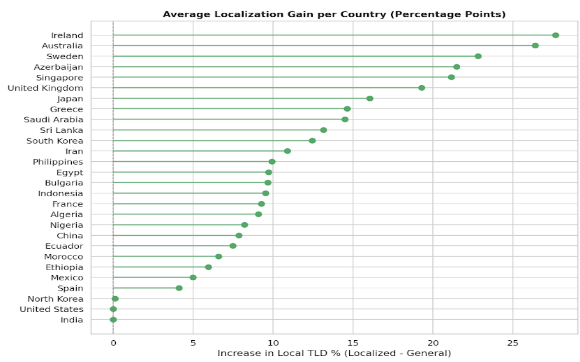

Figure 1: Average Localization Gain Per Country

4.3 Model Comparison

Although Qwen2.5-7B-Instruct and Aya-Expanse-8B achieve nearly identical overall accuracy at $k =$

Locale — Accuracy (%)											
am-ET	69.26	ar-DZ	86.46	ar-EG	83.42	ar-MA	72.56	ar-SA	71.17	as-AS	66.71
az-AZ	62.47	bg-BG	89.97	el-GR	81.27	en-AU	80.90	en-GB	81.26	en-US	88.98
es-EC	92.94	es-ES	85.19	es-MX	89.20	eu-PV	83.26	fa-IR	70.18	fr-FR	88.60
ga-IE	75.12	ha-NG	64.24	id-ID	72.63	ja-JP	77.56	ko-KP	65.81	ko-KR	87.30
su-JB	67.38	sv-SE	84.56	ta-LK	87.79	tl-PH	75.89	zh-CN	74.65	zh-SG	87.62
Overall: 78.81%											

Table 1: Per-locale accuracy of our submitted system (Qwen2.5-7B-Instruct, $k = 5$).

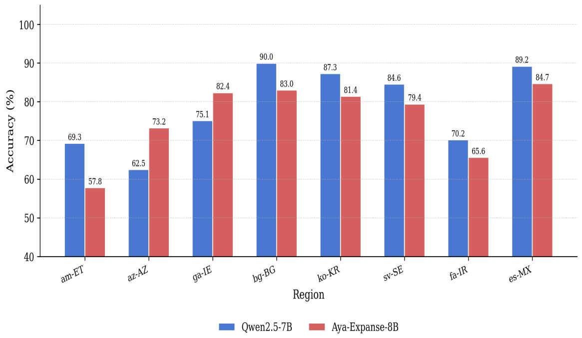

Figure 2: Per-locale accuracy of Qwen2.5-7B-Instruct and Aya-Expanse-8B at $k = 5$

5 (78.81% vs. 78.57%), per-locale performance reveals meaningful differences between the two models. Figure 2 shows the eight locales with the largest absolute differences. Since all questions are in English, these differences cannot be attributed to language understanding. They reflect differences in their ability to interpret and reason over culturally grounded evidence.

Qwen outperforms Aya on the majority of these locales, with particularly large margins on `am-ET` (+11.5%) and `bg-BG` (+7.0%). Aya, however, performs substantially better on `az-AZ` (+10.7%) and `ga-IE` (+7.3%). Notably, this does not follow cleanly from Aya's multilingual training focus, i.e., a model optimized for linguistic diversity does not automatically possess broader cultural knowledge. Cultural knowledge and linguistic knowledge appear to be dissociable in these models. Qwen's stronger performance on `am-ET` suggests its training data carries cultural coverage of Ethiopia that Aya's does not, despite Aya's explicit geographic diversity goals. The per-locale profile of each model is better understood as a fingerprint of its training data's cultural composition than of its multilingual capabilities.

4.4 Effect of Search Grounding

We examine how the number of retrieved search results affects performance by varying k, the number of search snippets prepended to the prompt, across $\{0, 1, 3, 5\}$. Setting $k = 0$ corresponds to the no-retrieval baseline, where the model relies entirely on its parametric knowledge. Table 2 reports overall accuracy for Qwen2.5-7B-Instruct and Aya-Expanse-8B at each value of k. Both models improve consistently as k increases, with Qwen rising from 72.40% at $k = 0$ to 78.81% at $k = 5$, and Aya from 73.04% to 78.57%. The gains are largest between $k = 0$ and $k = 1$, suggesting that even a single retrieved snippet provides a substantial signal, with diminishing but still positive returns at higher k.

Model	$k = 0$	$k = 1$	$k = 3$	$k = 5$
Qwen2.5-7B-Instruct	72.40	75.11	76.26	**78.81**
Aya-Expanse-8B	73.04	77.58	77.33	**78.57**

Table 2: Overall accuracy (%) by model and number of retrieved snippets k.

We additionally investigate how search grounding affects individual locales, shown in Figures 3 and 4. For lower-resource locales such as `am-ET`, `fa-IR`, and `sv-SE`, grounding with search results seems to yield clear gains in most cases. For example, `am-ET` improves from 60.1% to 69.3% and `sv-SE` from 74.7% to 84.6% when moving from $k = 0$ to $k = 5$. In contrast, locales that the model already handles well without retrieval tend to degrade slightly with added context. `en-US` drops from 92.7% to 89.0% and `zh-CN` from 79.6% to 74.7%, suggesting that for these locales the retrieved snippets introduce noise rather than useful signal, possibly because the model's parametric knowledge is already sufficient and the web context is redundant or misaligned. This pattern points to a broader finding - the benefit of search grounding is inversely related to how well-represented a locale is in the model's training data.

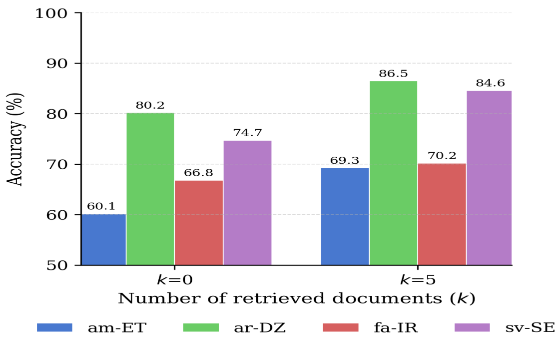

Figure 3: Locales where search grounding improves performance.

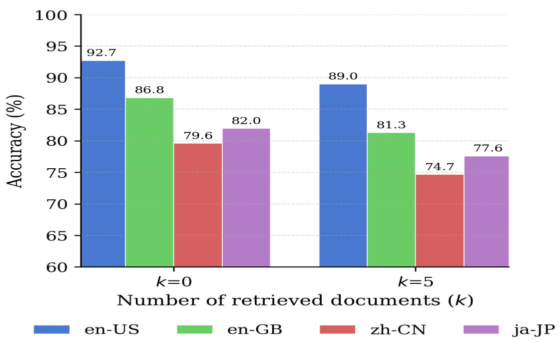

Figure 4: Locales where search grounding degrades performance.

4.5 Cultural Niche Framing

Inspired by the Thinking Through Other Minds (TTOM) framework (Veissiere et al., 2019), which models cultural cognition as inference within a shared cultural niche, we explored whether recasting the prompt in explicitly enculturated terms could improve performance. Rather than instructing the model to select the culturally specific answer, we reframed the task as reasoning from within the cultural niche of the target region. The revised prompt situates the model as an agent embedded in a particular cultural context, describes the distractor options as reflecting the regimes of attention of other cultural niches, and asks it to select the option that reflects what is salient or conventional specifically within the target region. The full revised prompt is included in Appendix D.

Table 3 reports the effect of this framing change at $k = 0$ and $k = 5$. At $k = 5$, the cultural niche framing yields consistent gains for both models: +2.23% for Qwen2.5-7B-Instruct (78.81% $\rightarrow$ 81.05%) and +1.42% for Aya-Expanse-8B (78.57% $\rightarrow$ 79.99%). At $k = 0$, however, the effect is negligible, with Qwen showing a marginal drop (−0.22%) and Aya a marginal gain (+0.54%). This interaction suggests that the cultural niche framing is most effective when grounded by retrieved context. Without external evidence, the more elaborate framing provides no additional signal for the model to reason over. Together, these results indicate that prompt framing and retrieval are complementary rather than independent levers for improving culturally grounded QA.

Model	Prompt	$k = 0$	$k = 5$
Qwen2.5-7B-Instruct	Baseline	72.40	78.81
	TTOM	72.18	**81.05**
Aya-Expanse-8B	Baseline	73.04	78.57
	TTOM	73.58	**79.99**

Table 3: Overall accuracy (%) with baseline and TTOM-inspired cultural niche prompts, at $k = 0$ and $k = 5$.

5 Conclusion

We presented a lightweight retrieval-augmented system based on locale-aware web search for culturally grounded QA. Our system requires no fine-tuning and builds entirely on compact open-source models. Locale-targeted web retrieval consistently improves performance, with the largest gains for lower-resource locales where parametric knowledge is weakest. We additionally demonstrate that localizing search parameters provably shifts the geographic composition of retrieved sources, and culturally informed prompt framing offers complementary gains when grounding context is available. Per-locale analysis reveals that models of similar overall accuracy can have starkly different cultural coverage profiles — a finding that has implications for how we evaluate and select models for culturally diverse deployments. More broadly, our results suggest that the web, when queried thoughtfully, remains a powerful and underutilized resource for bridging the cultural gaps of LLMs.

6 Limitations and Future Work

While locale-aware retrieval improves performance for lower-resource settings, our current pipeline applies retrieval uniformly across all locales, which can introduce noise for high-resource settings where parametric knowledge is already sufficient. This suggests the need for adaptive retrieval strategies that selectively invoke external grounding based on instance-level signals such as confidence or expected utility. Additionally, we use the question text directly as the search query, which keeps the pipeline simple but may introduce noise in MCQ settings due to the presence of distractor options; addressing this through query reformulation is left to future work. Our approach also assumes that retrieved web evidence is reliable, whereas real-world deployment requires handling noisy, adversarial, or conflicting sources through filtering, re-ranking, and uncertainty-aware aggregation. Our experiments are limited to compact models under realistic compute constraints; how retrieval-augmented cultural grounding scales with model size remains an interesting open direction as well. Finally, we plan to explore ensemble approaches that combine the complementary strengths of different models across regions.

References

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

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.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Piyawat Lertvittayakumjorn, David Kinney, Vinodkumar Prabhakaran, Donald Martin Jr., and Sunipa Dev. 2025. Towards geo-culturally grounded LLM generations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 313–330, Vienna, Austria. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

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.

Jean De Dieu Nyandwi, Yueqi Song, Simran Khanuja, and Graham Neubig. 2025. Grounding multilingual multimodal LLMs with cultural knowledge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24187–24231, Suzhou, China. Association for Computational Linguistics.

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.

Samuel Veissiere, Axel Constant, Maxwell Ramstead, Karl Friston, and Laurence Kirmayer. 2019. Ttom in action: Refining the variational approach to cognition and culture. *Behavioral and Brain Sciences*, 43.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

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 Serper.dev API Response Example

Below is an example raw JSON response returned by Serper.dev for the query *"Which city is the main destination for job seekers in Ireland?"* with `gl=ie` and `hl=ga`.

Listing 1: Example Serper.dev API response (truncated to 3 results).


```
{
  "searchParameters": {
    "q": "Which␣city␣is␣the␣main␣destination
␣␣␣␣␣␣␣␣␣␣for␣job␣seekers␣in␣Ireland?",
```

```
      "gl": "ie",
      "hl": "ga",
      "type": "search",
      "num": 5,
      "engine": "google"
    },
    "organic": [
      {
        "title": "Best␣Cities␣in␣Ireland␣for␣Job␣
        Seekers",
        "link": "https://jobvacanciesireland.com
        /...",
        "snippet": "Best␣Cities␣in␣Ireland␣for␣Job
        ␣Seekers;
␣␣␣␣␣␣␣␣Dublin,␣Tech,␣Finance,␣Consulting,␣High;
␣␣␣␣␣␣␣␣Cork,␣Pharma,␣Engineering,␣Tech,␣Medium;
␣␣␣␣␣␣␣␣Galway,␣Med-Tech,␣IT␣...",
        "position": 1
      },
      {
        "title": "Exploring␣the␣Job␣Market␣in␣
        Ireland
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣[2024-2025]",
        "link": "https://reliancerecruitment.ie/...
        ",
        "snippet": "The␣country␣is␣home␣to␣several
␣␣␣␣␣␣␣␣multinational␣corporations␣and␣startups,
␣␣␣␣␣␣␣␣particularly␣in␣cities␣like␣Dublin,
␣␣␣␣␣␣␣␣Cork,␣and␣Galway.",
        "position": 2
      },
      {
        "title": "Dublin␣and␣Antrim␣show␣highest␣
        job
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣vacancy␣rates␣in␣Ireland",
        "link": "https://www.irishjobs.ie/...",
        "snippet": "Dublin's␣high␣vacancy␣rate␣
        (4.96)␣and
␣␣␣␣␣␣␣␣its␣high␣employment␣rate␣put␣it␣ahead
␣␣␣␣␣␣␣␣of␣any␣county␣in␣the␣Republic
␣␣␣␣␣␣␣␣of␣Ireland.",
        "position": 3
      }
    ],
    "credits": 1
}
```


This is then formatted into the following context block prepended to the model prompt:

Listing 2: Formatted search context block passed to the model.

```
Search results:

1. Best Cities in Ireland for Job Seekers
   Best Cities in Ireland for Job Seekers;
   Dublin, Tech, Finance, Consulting, High;
   Cork, Pharma, Engineering, Tech, Medium;
   Galway, Med-Tech, IT ...

2. Exploring the Job Market in Ireland
    [2024-2025]
   The country is home to several multinational
   corporations and startups, particularly in
   cities like Dublin, Cork, and Galway.

3. Dublin and Antrim show highest job vacancy
   rates in Ireland
   Dublin's high vacancy rate (4.96) and its
   high employment rate put it ahead of any
   county in the Republic of Ireland.
```

B Prompt Template

Below is the prompt template used for inference. The [SEARCH CONTEXT] block is included only when $k > 0$; it is omitted entirely for the no-retrieval baseline.

Listing 3: Full prompt template.

```
You are answering a cultural knowledge
question. Select the option that is
SPECIFIC to the region/culture mentioned
in the question.

Important: The incorrect options are from
 regions/cultures.

Given below are the top search results
that come up upon googling the given
question. You may use them if relevant:

[SEARCH CONTEXT]

Question: <question text>

A) <option A>
B) <option B>
C) <option C>
D) <option D>

Answer with ONLY the letter (A, B, C, or D):
```

C TLD Localization Gain

Below is an example of how the search results vary with a localization parameter being passed. Localized search results pick more regional domains.

Listing 4: General Result

```
Question: From what age is an individual allowed
      to work in Sweden? (Provide Arabic numerals
      (e.g., 12) only.)

URLs hit:

https://www.worldvaluessurvey.org/WVSContents.
     jsp?CMSID=FieldworkSampling
http://www.mdcourts.gov/circuit/howard/
     juryservice
https://en.wikipedia.org/wiki/
     National_identification_number
```

Listing 5: Localised Results

```
Question: From what age is an individual allowed
      to work in Sweden? (Provide Arabic numerals
      (e.g., 12) only.)

URLs hit:
```

```
https://www.skatteverket.se/servicelankar/
    otherlanguages/englishengelska/
    individualsandemployees/movingtosweden/
    citizenofeueeacountry/youaregoingtowork.4.5
    a85666214dbad743ffff42.html
https://www.uhr.se/en/start/recognition-of-
    foreign-qualifications/enic-naric-sweden/
    qualification-frameworks-for-swedish-
    qualifications-and-degrees/
https://www.informationsverige.se/en/om-sverige/
    att-forsorja-sig-och-utvecklas-i-sverige/att
    -vara-anstalld.html
```

D TTOM-Inspired Prompt Template

Below is the revised prompt used for the cultural niche framing experiments.

Listing 6: TTOM-inspired cultural niche prompt template.

```
You are answering a cultural knowledge
question from within the cultural niche
of <region name>.

Within this cultural context, certain
things carry meaning and salience that
would be invisible or irrelevant to
outsiders.

The incorrect options reflect the regimes
of attention of OTHER cultural niches --
they are correct answers for different
cultures, not this one.

Given below are the top search results
that come up upon googling the given
question. You may use them if relevant:

[SEARCH CONTEXT]

Select the option that reflects what is
salient, meaningful, or conventional
specifically within <region name>.

Question: <question text>

A) <option A>
B) <option B>
C) <option C>
D) <option D>

Answer with ONLY the letter (A, B, C, or D):
```