LocuPrompt at SemEval-2026 Task 7: A Multilingual Prompting Framework for Cross-Cultural Everyday Knowledge in LLMs

Ning Jingke
Yunnan University, Kunming, China
20221060109@mail.ynu.edu.cn

Abstract

Understanding everyday cultural knowledge remains a fundamental challenge for large language models (LLMs). This paper presents LOCUPROMPT, a multilingual framework for SemEval-2026 Task 7: *Everyday Knowledge Across Diverse Languages and Cultures*. To address Short Answer Questions (SAQ), we employ an English-pivot generation strategy with back-translation, combined with empirical locale-specific routing that dynamically assigns the optimal LLM to each target region. For Multiple-Choice Questions (MCQ), we apply parameter-efficient fine-tuning to a robust multilingual base model, utilizing locale-aware instructions that frame the LLM as a "local resident." By integrating strategic model selection with resource-efficient adaptation, LOCUPROMPT effectively bridges cross-lingual cultural gaps while maintaining a fully reproducible pipeline.

1 Introduction

Large language models (LLMs) excel at general knowledge and logical reasoning, but their understanding of everyday cultural knowledge—the implicit, context-sensitive wisdom that guides daily life in specific societies—remains fragile and unevenly distributed. Questions like "What do people in Morocco say when someone sneezes?" or "Which hand is considered polite to eat with in Saudi Arabia?" cannot be answered by world knowledge alone; they require deep familiarity with local customs, values, and lived experience. SemEval-2026 Task 7 introduces extended BLEnD(Myung et al., 2024), a multilingual benchmark to evaluate this capability across over 30 language-region pairs, spanning six continents and diverse cultural spheres.

To address this challenge, this paper proposes LOCUPROMPT, a lightweight, prompting-based framework that adapts to cultural context without extensive training. LOCUPROMPT tackles both task tracks: (1) Short Answer Questions (SAQ), where a pivot-language pipeline combines culturally grounded prompting with optional back-translation; and (2) Multiple-Choice Questions (MCQ), where a pre-trained multilingual model is adapted via parameter-efficient fine-tuning on the provided development data, using locale-aware instructions that frame the model as a "local resident."

A key design principle of LOCUPROMPT is empirical model selection. Rather than relying on a single LLM, the framework assigns the best-performing model—selected from a diverse pool of state-of-the-art systems—to each locale based on validation performance. This data-driven routing avoids the "one-model-fits-all" pitfall and leverages the complementary cultural strengths of different models.

This work demonstrates that combining culturally contextualized prompting, strategic model selection, and minimal supervised adaptation can effectively enhance cross-cultural performance in multilingual question answering. All methodological components are designed for reproducibility under the constraints of the shared task.

2 Related Work and Task Background

2.1 Knowledge and Cultural Evaluation

SemEval-2026 Task 7 (Ousidhoum et al., 2026) addresses a critical gap in LLM evaluation by introducing the **extended BLEnD** benchmark. This task evaluates LLMs' understanding of region-specific common sense, ranging from local customs to practical social norms, through two formats: Short Answer Questions (SAQ) and Multiple-Choice Questions (MCQ).

This benchmark transcends traditional multilingual datasets like MKQA (Longpre et al., 2021) and XQuAD (Schuster et al., 2019), which fo-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 973–980

cus heavily on the cross-lingual transfer of *factual* knowledge while overlooking *culture-specific* nuances. While XCOPA (Ponti et al., 2020) explores causal commonsense across cultures, its scenarios are often abstract and decoupled from specific geographical locales. Furthermore, recent studies on cultural bias often rely on high-level national indices, such as Hofstede scores (Pawar et al., 2024); however, these indices fail to capture the granular, lived experiences that characterize everyday cultural competence.

2.2 Prompting and Cultural Alignment

Our methodology builds on advances in prompting strategies and model selection. In terms of prompting, techniques such as Chain-of-Thought (CoT)(Wei et al., 2023; Kojima et al., 2023) have demonstrated efficacy in decomposing complex reasoning tasks. To bridge the cultural gap, role-play prompting has emerged as a powerful tool for cultural alignment(Kong et al., 2024), alongside broader efforts in culturally aware prompting to mitigate socio-cultural biases (Adilazuarda et al., 2024).

2.3 Model Selection and Routing

The concept of routing queries to specialized systems—a core feature of LocuPrompt—draws inspiration from the Mixture-of-Experts (MoE) paradigm (Shazeer et al., 2017; Fedus et al., 2022) and ensemble selection frameworks like LLM-Blender (Jiang et al., 2023). While MoE focuses on architectural efficiency, our approach applies the principle of empirical model selection to leverage the idiosyncratic cultural strengths of different LLMs. Despite the success of these techniques in general domains, their application to diverse and context-heavy locales, locale-specific adaptation remains underexplored. SemEval-2026 Task 7 provides a unique testbed for solutions that are both computationally accessible and culturally grounded, exposing existing limitations in models' ability to navigate diverse real-world social landscapes.

3 Methodology

LocuPrompt is a modular framework designed for SemEval-2026 Task 7, integrating culturally grounded prompting, empirical model selection, and parameter-efficient fine-tuning (PEFT).

3.1 Short Answer Questions (SAQ)

For the SAQ track, we implement a **pivot-language prompting pipeline** that leverages the reasoning capabilities of English-dominant LLMs while preserving cultural context. The pipeline consists of five stages:

1. **Input**: A question q_ℓ in the target language ℓ (e.g., Irish: *"Cé na tionscail a roghnaíonn daoine óga..."*).
2. **Translation**: q_ℓ is translated into English (q_{en}) via the **Google Translate API**. This avoids the instability of zero-shot in-context translation in multilingual LLMs.
3. **Culturally Grounded Prompting**: q_{en} is embedded into a role-play prompt framing the model as a "local resident."

   ```
   ### Instruction: You are a
   local resident of [Country].
   Answer the following question
   in English, concisely and with
   cultural accuracy. Provide
   only the essential answer
   without preamble.
   ### Question: q_en
   ### Response:
   ```

4. **Generation**: The selected LLM generates a concise English answer a_{en}.
5. **Back-translation**: a_{en} is translated back into language ℓ to produce the final answer a_ℓ.

3.2 Multiple-Choice Questions (MCQ)

For the MCQ track, we employ Quantized Low-Rank Adaptation (QLoRA) (Hu et al., 2021; Dettmers et al., 2023) to adapt a strong multilingual model to the nuances of the task. Given the relatively small size of the provided pilot data, QLoRA allows for efficient supervised fine-tuning (SFT) by updating only a small fraction of the model's parameters.

Model Configuration: We utilize a 4-bit quantized base model (e.g., Qwen series) as the backbone. LoRA adapters are applied to all linear layers (including q, k, v, o projections and MLP layers) with a rank of $r = 16$ and $\alpha = 16$ to ensure sufficient capacity for learning cultural contexts.

Task Framing: Each training sample is converted into a structured instruction following the prompt template:

```
"As a local resident of
[Country], please answer the
following question based on
```

common knowledge in [Country].
Question: q
Options: A. o_1 B. o_2 C. o_3 D. o_4
Answer with only the letter (A, B, C, or D):"

Optimization: The model is trained to minimize the negative log-likelihood of the correct answer letter y:

$$\mathcal{L} = -\log P(y \mid X; \theta_{\text{frozen}} + \Delta\theta_{\text{LoRA}})$$

where X is the formatted input and $\Delta\theta_{\text{LoRA}}$ represents the trainable low-rank matrices. This strategy allows the model to align its outputs with the specific MCQ format and internalize region-specific social norms provided in the development set.

3.3 Locale-Specific Model Selection

A primary innovation of LocuPrompt is its **empirical model routing** mechanism. Recognizing that no single LLM dominates across all cultural contexts—with performance gaps exceeding 35% accuracy in certain locales—we dynamically assign the most proficient model to each language-region pair.

Formally, let $\mathcal{M} = \{m_1, \ldots, m_n\}$ be the pool of candidate models, and $s(c, m)$ represent the validation accuracy for a pilot locale $c \in \mathcal{C}_{\text{pilot}}$. The optimal model $m^*(c)$ is defined as:

$$m^*(c) = \arg\max_{m \in \mathcal{M}} s(c, m)$$

For test locales $c_{\text{test}} \notin \mathcal{C}_{\text{pilot}}$, we employ a **linguistic and geographic proximity heuristic** to map unseen regions to the most proximate pilot locale. For example, `en-US` inherits the model selected for `en-GB`, and `zh-TW` defaults to the choice for `zh-CN`. While we acknowledge that nearby cultures exhibit meaningful local variations (Bromham and Yaxley, 2023), this strategy leverages the correlation between linguistic families and regional contexts to generalize model performance (Koto et al., 2024).

Limitation. We acknowledge that this proximity-based mapping is an approximation. As noted in cross-cultural studies, nearby cultures share many traits but also exhibit meaningful local variation (Bromham and Yaxley, 2023). For instance, while `en-US` and `en-GB` share a language, everyday cultural knowledge (e.g., social norms, customary practices) can differ substantially. Future work should collect pilot data for underrepresented locales to enable direct empirical selection rather than relying on proximity heuristics. Despite this limitation, the routing strategy significantly outperforms any single model, as evidenced by the large performance gaps observed in pilot locales.

4 Experimental Setup

4.1 Data Splits and Usage

LocuPrompt strictly adheres to the competition's data constraints, using no external training data. The official dataset consists of:

- **Pilot (Trial) Data**: ∼148 examples across 24 language-region pairs, used as our development set.
- **Test Data**: A hidden set of 31 locales, including several not present in the pilot set.

For Track 1 (SAQ), the pipeline is entirely zero-shot; pilot data is used only for model selection per locale. For Track 2 (MCQ), we perform supervised fine-tuning (SFT) using the pilot data.Given the limited sample size (($N \approx 6$ per locale), **5-fold cross-validation** was conducted on the pilot data to obtain a more **statistically rigorous estimate** of the fine-tuning performance.

4.2 System Design Rationale

During development, we compared several architectural variants to optimize the balance between cultural nuance and reasoning robustness:

- **Uniform Model Baseline**: A high-capacity multilingual model (one-size-fits-all) used to establish a performance floor.
- **Direct Target-Language Generation**: Prompts were executed directly in the target language. This led to frequent hallucinations and degraded fluency in morphologically complex or low-resource languages like Irish (`ga-IE`) and Basque (`eu-ES`).
- **English-Pivot Pipeline**: Generating responses in English before translating back to the target language.

Justification for the English-Pivot Strategy: While pivoting may introduce translation artifacts, our pilot experiments confirmed its systemic superiority. The English-pivot strategy achieved an overall accuracy of **54.73%**, significantly outperforming direct generation (**45.95%**). This 8.8-point margin suggests that the reasoning depth acquired during English-centric pre-training outweighs the

potential information loss during back-translation. This strategy ensures logical consistency across diverse locales while our model routing mechanism further refines cultural sensitivity at the individual model level.

4.3 Training Details (Track 2)

Model selection. We choose Qwen3-4B (Team, 2025) as our base model due to its strong multilingual capabilities and favorable performance-to-size ratio. This decision is primarily motivated by resource constraints common to individual researchers: the model fits within a single 24GB GPU and can be fine-tuned without exceeding memory limits. Alternative models such as LLaMA-3-8B present two barriers: first, they require an official request and approval from Meta, which introduces administrative overhead for individual researchers; second, with 8 billion parameters, LoRA-based adaptation would risk exceeding our 24GB memory budget, especially when accommodating batch processing and gradient states. Larger models such as Mixtral-8×7B (effectively 47B parameters with MoE) are prohibitively memory-intensive under our hardware constraints. Our goal is not to propose a state-of-the-art system, but to establish a reproducible and lightweight baseline that lower-resource participants can easily build upon.

Hyperparameter	Value
LoRA Rank (r) / Alpha (α)	16 / 16
Learning Rate	2×10^{-4} (Constant)
Optimizer	8-bit AdamW
Batch Size	4 (Accumulation = 2)
Epochs	3
Max Sequence Length	512 tokens

Table 1: Hyperparameters for MCQ fine-tuning.

QLoRA adapters are applied to all linear layers (`q, k, v, o, gate, up, down_proj`). Training was conducted on a single 24GB GPU, requiring approximately 10 minutes per locale. Given the small-scale data, we emphasize that SFT serves to adapt the model's output format rather than injecting broad knowledge.

Observations on SFT Performance. Preliminary findings from our cross-validation suggest that supervised fine-tuning on such limited pilot data is exceptionally challenging, with performance remaining near the random baseline. This underscores a key characteristic of the task: nuanced cultural common sense cannot be readily "injected" into a model through ultra-low-resource SFT alone. These empirical results further validate our design choice to prioritize empirical model routing as more robust mechanisms for cross-cultural adaptation.

4.4 External Tools and Libraries

The framework is established using *Hugging Face Transformers* and *PEFT*. Translation is handled by a commercial machine translation API. We evaluated open-weight models and commercial APIs with temperature settings $\in [0.1, 1.1]$. Our solution is designed for reproducibility and serves as a lightweight baseline for lower-resource participants.

5 Results and Analysis

5.1 Overall Performance

Tables 2 and 3 present the exact-match (EM) accuracy of LocuPrompt on the official test set for Track 1 (SAQ) and Track 2 (MCQ), respectively.

Track 1 (SAQ). Our system achieves an overall accuracy of **52.14%**, ranking 6th out of 13 participating systems. Performance varies substantially across locales, ranging from 13.4% (ar-MA) to 77.6% (en-US). Notably, English locales consistently outperform their non-English counterparts (e.g., en-EG: 59.6% vs. ar-EG: 48.0%; en-SA: 50.8% vs. ar-SA: 40.0%), reflecting the challenge of multilingual generation for SAQ. The highest-performing locale is en-US (77.6%), while the lowest is ar-MA (13.4%), suggesting that North African Arabic dialects pose particular difficulty for zero-shot generation.

Track 2 (MCQ): LocuPrompt attained an overall accuracy of **76.47%**, ranking 13th out of 19. Accuracy was generally higher than in SAQ due to the constrained nature of the task. The highest scores were observed in `es-EC` (92.94%) and `en-US` (90.63%), while `ko-KP` (55.51%) remained the most challenging. Interestingly, the gap between English and target languages was less pronounced in MCQ, suggesting that SFT on pilot data effectively helped bridge linguistic barriers.

5.2 Performance by Region and Language Family

To identify systematic patterns, we aggregated results by geographic and linguistic categories.

Regional Disparities: In SAQ, North African locales (`ar-MA`, `ar-DZ`, `ar-EG`) showed the lowest average accuracy (35.7%), likely due to the divergence of local dialects from Modern Standard Ara-

Table 2: Track 1 (SAQ) results by locale: exact-match accuracy (%). Overall accuracy: 52.14%. Rank: 6/13.

Locale	Acc. (%)	Locale	Acc. (%)
am-ET	37.2	en-ET	41.6
ar-DZ	45.8	en-DZ	64.8
ar-EG	48.0	en-EG	59.6
ar-MA	13.4	en-MA	24.4
ar-SA	40.0	en-SA	50.8
as-AS	32.4	en-AS	56.2
az-AZ	38.8	en-AZ	41.8
bg-BG	34.6	en-BG	35.0
el-GR	40.0	en-GR	43.6
en-AU	70.2	en-GB	61.6
en-US	77.6	en-SG	74.0
es-EC	54.6	en-EC	52.4
es-ES	71.8	en-ES	65.2
es-MX	60.2	en-MX	55.2
eu-PV	54.6	en-PV	53.0
fa-IR	71.2	en-IR	65.8
fr-FR	64.8	en-FR	59.2
ga-IE	48.8	en-IE	52.0
ha-NG	34.4	en-NG	31.8
id-ID	76.6	en-ID	72.6
ja-JP	54.8	en-JP	51.6
ko-KP	48.4	en-KP	53.6
ko-KR	72.8	en-KR	73.4
ms-SG	64.2	zh-TW	29.2
su-JB	57.2	en-JB	56.0
sv-SE	40.0	en-SE	37.0
ta-LK	26.4	en-LK	31.4
ta-SG	47.2	tl-PH	56.4
zh-CN	66.8	en-CN	69.4
zh-SG	64.6	en-PH	53.6
en-TW	51.2		

Table 3: Track 2 (MCQ) results by locale: exact-match accuracy (%). Overall accuracy: 76.47%. Rank: 13/19.

Locale	Acc. (%)	Locale	Acc. (%)
am-ET	57.21	ar-DZ	81.31
ar-EG	73.91	ar-MA	67.59
ar-SA	74.77	as-AS	65.85
az-AZ	63.43	bg-BG	75.31
el-GR	84.09	en-AU	86.94
en-GB	82.10	en-US	90.63
es-EC	92.94	es-ES	84.00
es-MX	89.36	eu-PV	76.37
fa-IR	62.34	fr-FR	86.32
ga-IE	80.72	ha-NG	58.81
id-ID	72.08	ja-JP	84.15
ko-KP	55.51	ko-KR	86.07
su-JB	70.87	sv-SE	71.36
ta-LK	81.69	tl-PH	70.16
zh-CN	87.20	zh-SG	81.07

bic. In contrast, East Asian locales (`zh`, `ja`, `ko`) achieved the highest performance (average 66.2%). In MCQ, regional gaps narrowed, though the Korean Peninsula remained a difficult outlier (average 70.8%).

Language Family: English locales outperformed non-English locales in SAQ (59.4% vs. 44.9%). However, in MCQ, this gap nearly disappeared (78.2% vs. 75.1%), indicating that while multilingual generation is brittle, multilingual *comprehension* and selection are more robust after adaptation.

5.3 Ablation Study and Cross-Validation

Given the small pilot set ($N = 148$), we performed ablation tests to evaluate our design choices.

- **Model Selection:** Implementing **empirical model routing** per locale yielded a 2.7% improvement over a fixed-model baseline (increasing from 52.0% to 54.7% on pilot data).
- **SFT vs. Prompting:** For MCQ, LoRA finetuning provided a marginal 1.2-point gain over the zero-shot baseline. However, 5-fold cross-validation revealed an average accuracy of only 27.13% ($\pm$7.16%) on the pilot set. This near-baseline performance confirms that 148 examples are insufficient to "inject" new cultural knowledge; SFT primarily serves to align the model with the output format.
- **Translation Impact:** The English-pivot strategy outperformed direct generation by 8.8 points

(54.73% vs 45.95%). However, this benefit is uneven: in low-resource locales like `ga-IE` and `ar-MA`, translation noise significantly degraded SAQ accuracy compared to their English equivalents.

5.4 Error Analysis

We conducted a quantitative error analysis on a sampled subset of 200 cases (100 per track) to identify recurring failure modes.

Failure Modes in SAQ:

- **Factual Inaccuracy (87%):** The primary error source, where the model lacked specific regional knowledge (e.g., local ingredients or startup trends).
- **Translation Artifacts (10%):** In languages like Irish and Basque, the pipeline misinterpreted dialectal terms, leading to nonsensical responses.
- **Cultural Generalization (1%):** The model occasionally defaulted to global stereotypes (e.g., "bowing" for all Asian greetings) rather than specific local practices.

Failure Modes in MCQ:

- **Factual Errors (82%):** Similar to SAQ, the model often selected incorrect options due to a lack of localized common sense.
- **Distractor Confusion (17%):** The model struggled with fine-grained cultural differences between semantically similar options (e.g., choosing a "nod" instead of a "bow").
- **Question Ambiguity (1%):** Some questions lacked sufficient context, allowing for multiple plausible interpretations.

6 Conclusion

This work introduced LocuPrompt, a lightweight framework for cultural QA that combines empirical model selection, English-pivot prompting, and parameter-efficient adaptation. Our results on the SemEval-2026 Task 7 benchmark demonstrate that:

1. **Locale-aware model routing** effectively leverages the unevenly distributed cultural strengths of different LLMs.
2. **English-pivot pipelines** significantly extend the reasoning depth for most locales but remain fragile for low-resource languages due to translation bottlenecks.
3. **Small-scale SFT** is effective for format alignment but cannot compensate for a baseline lack of localized factual knowledge.

Future work should focus on grounding models in regionally curated data to mitigate factual errors and exploring few-shot in-context learning to navigate data-scarce cultural regimes.

A Model Selection and Additional Details

A Per-Locale Model Performance

Table 4 reports the exact-match accuracy (%) of five anonymized multilingual language models on the official pilot data across 24 language-region pairs. For each locale, all models achieving the highest score are bolded.

To guide model selection for the test submission, we adopted a pilot-based routing strategy: for each locale present in the pilot set, we selected the model with the highest validation performance. For locales absent from the pilot (e.g., en-US, zh-TW), we assigned models based on linguistic or geographic proximity to a pilot locale (e.g., en-US $\rightarrow$ en-GB, zh-TW $\rightarrow$ zh-CN). The final mapping is provided in Table 5.

Locale	Lamma3	Qwen	DeepSeek	GPT-5	Gemini
Overall	39.2	43.9	45.9	45.9	**52.0**
ar-EG	28.6	28.6	42.9	42.9	**57.1**
ar-MA	**57.1**	**57.1**	**57.1**	**57.1**	**57.1**
ar-SA	42.9	42.9	**57.1**	**57.1**	**57.1**
bg-BG	**71.4**	57.1	57.1	57.1	57.1
el-GR	**60.0**	**60.0**	**60.0**	**60.0**	**60.0**
en-AU	28.6	14.3	28.6	**42.9**	**42.9**
en-GB	**60.0**	20.0	20.0	40.0	40.0
es-EC	62.5	75.0	**87.5**	62.5	75.0
es-ES	**60.0**	**60.0**	**60.0**	**60.0**	**60.0**
es-MX	20.0	**40.0**	20.0	20.0	**40.0**
eu-ES	14.3	42.9	42.9	57.1	**71.4**
fa-IR	40.0	40.0	40.0	40.0	**60.0**
fr-FR	0.0	0.0	**25.0**	12.5	**25.0**
ga-IE	14.3	**28.6**	0.0	14.3	14.3
id-ID	20.0	40.0	40.0	**60.0**	40.0
ja-JP	**14.3**	**14.3**	**14.3**	**14.3**	**14.3**
ko-KR	0.0	0.0	20.0	20.0	**40.0**
ms-SG	57.1	**71.4**	57.1	57.1	**71.4**
ta-LK	**71.4**	57.1	57.1	57.1	**71.4**
ta-SG	**28.6**	**28.6**	**28.6**	14.3	14.3
tl-PH	50.0	75.0	**87.5**	75.0	75.0
zh-CN	60.0	**100.0**	80.0	**100.0**	**100.0**
zh-SG	42.9	**57.1**	**57.1**	42.9	**57.1**

Table 4: Per-locale exact-match accuracy (%) of five models on the pilot data. All highest scores per row are bolded.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and

Table 5: Mapping from country code to the selected best-performing model for test submission. For locales absent from the pilot data (e.g., US, TW), the assignment is based on linguistic proximity to a pilot locale (e.g., en-US → en-GB, zh-TW → zh-CN). Model identifiers are anonymized for double-blind review.

Country Code	Country / Region	Selected Model
EG	Egypt	Gemini
MA	Morocco	Lamma3
SA	Saudi Arabia	DeepSeek
BG	Bulgaria	Lamma3
GR	Greece	Lamma3
AU	Australia	Gemini
GB	United Kingdom	Lamma3
EC	Ecuador	DeepSeek
ES	Spain	Gemini
MX	Mexico	Qwen
IR	Iran	Gemini
FR	France	DeepSeek
IE	Ireland	Qwen
ID	Indonesia	GPT-5
JP	Japan	Qwen
KR	South Korea	Gemini
SG	Singapore	Qwen
LK	Sri Lanka	Lamma3
PH	Philippines	DeepSeek
CN	China	Qwen
ET	Ethiopia	Gemini
DZ	Algeria	Gemini
NG	Nigeria	DeepSeek
AS	India (Assam)	GPT-5
AZ	Azerbaijan	Lamma3
US	United States	GPT-5
PV	Basque Country (Spain)	Gemini
JB	Indonesia (West Java)	GPT-5
SE	Sweden	Lamma3
KP	North Korea	Gemini
TW	Taiwan	Qwen

modeling "culture" in llms: A survey. *Preprint*, arXiv:2403.15412.

Lindell Bromham and Keaghan J. Yaxley. 2023. Neighbours and relatives: accounting for spatial distribution when testing causal hypotheses in cultural evolution. *Evolutionary Human Sciences*, 5:e27.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Preprint*, arXiv:2101.03961.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. In *arXiv preprint*. arXiv.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. *Preprint*, arXiv:2306.02561.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. In *arXiv preprint*. arXiv.

Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024. Indoculture: Exploring geographically-influenced cultural commonsense reasoning across eleven indonesian provinces. *Preprint*, arXiv:2404.01854.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. In *arXiv preprint*. arXiv.

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.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. Survey of cultural awareness in language models: Text and beyond. In *arXiv preprint*. arXiv.

Edoardo M. Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *arXiv preprint*. arXiv.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *Preprint*, arXiv:1701.06538.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.