Multilingual Question Answering Considering Language- and Culture-Specific Knowledge

Xichen ZHANG, Satoshi KOSUGI, Kotaro FUNAKOSHI, Manabu OKUMURA
Institute of Science Tokyo
`{zhang_x,kosugi,funakoshi,oku}@lr.first.iir.isct.ac.jp`

Abstract

Large language models (LLMs) often underperform in non-English settings, where failures can arise from misinterpretations triggered by missed or misread language- and culture-specific cues rather than limitations in reasoning alone. To address this, we propose **Language- and Culture-Aware Prompting (LCAP)**, a two-stage framework that first extracts cues and explains the induced meaning shifts, then constrains answer selection to the extracted evidence. On mCSQA (`hard`) [1] across seven non-English languages and three LLMs, LCAP consistently outperforms both direct prompting and cross-lingual prompting (CLP), achieving at least a 2.3 points average gain over the baseline.

1 Introduction

Large language models (LLMs) are widely used in natural language processing (NLP) applications [2]. Prompting strategies such as Chain-of-Thought (CoT) can significantly improve performance on reasoning tasks by encouraging step-by-step reasoning [3]. However, because training and evaluation pipelines are often English-centered, LLMs may underperform or exhibit English-centric biases in non-English settings, limiting their global applicability [4, 5]. In addition, recent work reports that LLMs may also favor English or Western cultural values even when prompted in Arabic settings, raising concerns of misunderstanding and miscommunication in cross-cultural interactions [6].

This issue in multilingual settings is not solely due to limitations in LLMs' reasoning. Evidence from multilingual transformers shows that cross-lingual representation alignment correlates strongly with transfer performance, suggesting that misalignment can contribute to non-English errors [7]. Recent work further argues that many English-centric LLMs rely on an implicit English-pivot mechanism when processing non-English inputs, and that insufficient supervision for intermediate-layer alignment can yield misaligned representations and limit non-English performance [8].

Prior studies have explored various prompting strategies. AlKhamissi et al. [9] ask the model to answer as a real-life person with a particular background. Culturally-Grounded Chain-of-Thought (CG-CoT) constructs an external database of similar Yoruba and English proverbs, and prompts the model to contrast culturally grounded meanings [10]. Cross-lingual prompting (CLP) [11], a closely related direction to our work, generalizes CoT across languages by using a two-stage structure that separates question alignment from task solving. However, since the benchmarks in this line of work are mainly elementary-school-level math problems, it remains unclear whether such methods can improve LLMs' performance on commonsense, language- and culture-specific questions.

To address this gap, we propose **Language- and Culture-Aware Prompting (LCAP)**, a two-stage prompting framework that (*i*) explicitly extracts language-/culture-specific cues and explains their induced meaning shifts, and (*ii*) constrains answer selection to the extracted evidence. We conducted experiments on three representative LLMs: GPT-3.5-Turbo, GPT-5-Nano, and Gemini-2.5-Flash-Lite. Across the models, LCAP consistently outperforms both Direct, prompting the model to answer the original question directly, and CLP, achieving more than 2.3 points average gain.

2 Related Work

This section reviews Chain-of-Thought (CoT) prompting [3] and cross-lingual prompting (CLP) [11]. We also discuss why multilingual question answering may require

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

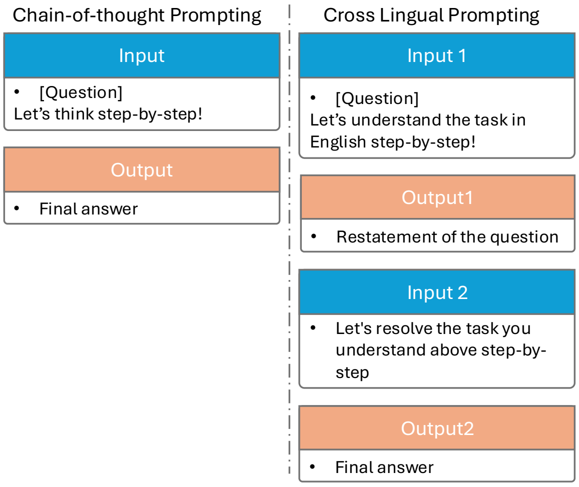

Chain-of-thought Prompting
Input
[Question]
Let's think step-by-step!
Output
Final answer
Cross Lingual Prompting
Input 1
[Question]
Let's understand the task in English step-by-step!
Output1
Restatement of the question
Input 2
Let's resolve the task you understand above step-by-step
Output2
Final answer

Figure 1 Overview of chain-of-thought [3] and cross-lingual prompting [11]

additional constraints beyond a generic alignment step.

Chain-of-Thought: CoT improves LLM performance by encouraging intermediate reasoning steps, often via a zero-shot trigger such as "Let's think step by step!" (Figure 1, left).

Cross-lingual Prompting: Cross-lingual CoT is commonly instantiated by prompting intermediate reasoning in English for non-English inputs, a setup widely used in multilingual math benchmarks such as MGSM [12]. CLP generalizes this idea with a two-stage design: a **cross-lingual alignment stage** (constructing an English understanding of the original question) followed by a **task solving stage** based on that aligned understanding (Figure 1, right).

Why Additional Constraints: Despite its effectiveness, CLP primarily treats alignment as a restatement step, which may overlook language/culture-specific cues in the original context. Even when an English restatement is fluent, it may fail to preserve the intended meaning if decisive cues are implicit or culturally grounded. In such cases, the failure occurs at the interpretation stage, and downstream reasoning can remain internally consistent while still producing an incorrect answer. This motivates our hypothesis that explicitly surfacing language/culture-specific cues may further strengthen LLMs' ability on multilingual question answering tasks.

3 LCAP: Language- and Culture-Aware Prompting

In this section, we present **Language- and Culture-Aware Prompting (LCAP)**, a two-stage prompting framework for language- and culture-specific questions designed to make the intermediate evidence explicit and enforceable.

3.1 Problem Setup

In this work, we focus on **language- and culture-specific questions**. To operationalize this focus, we consider an English native speaker and a non-English native speaker, a question is language- and culture-specific if, when the two speakers are given the original question and its machine-translated English version, respectively, either (1) **answer divergence**: the two speakers produce different answers, or (2) **unanswerability**: the English speaker fails to provide a meaningful answer. Here we introduce two illustrative examples of language- and culture-specific questions, the second one is from [1].

Case (1): Consider the question: "When you invite him to dinner, he smiles and shakes his head, which means: Agree or Reject?" When asked the English and Hindi version of this same question repectively, an English native speaker may choose Reject (head shake = no), while the Hindi speaker will likely choose Agree (head shake can signal affirmation) leading to **answer divergence**.

Case (2): Next, consider a Japanese multiple-choice question asking which of the following first-person pronouns sounds older: わし, わたし, おれ, ぼく, うち. A literal machine translation may render these options as the same English form (e.g., "I" or "me"), collapsing distinctions that are crucial for answering the question and thus making it effectively **unanswerable**

3.2 Stage 1: Cue Extraction and Meaning Explanation

Stage 1 targets failures that occur before reasoning begins, where decisive cues may be implicit and thus may be lost under a generic English restatement. Instead of paraphrasing the entire question, LCAP prompts the model to extract language-/culture-specific cues in the source language. In the multiple-choice setting, this extraction is performed for both the question stem and each answer option, using a fixed format (`Cue` → `Meaning shift`). The model then explains the extracted meaning shifts in English step by step, without providing a full translation. The resulting evidence is later used as the sole basis for decision making.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

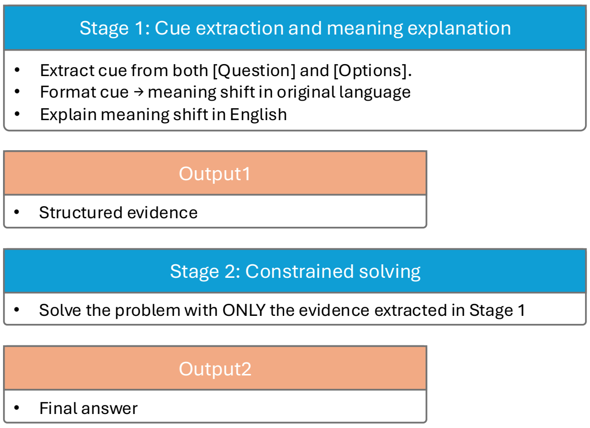

Stage 1: Cue extraction and meaning explanation
Extract cue from both [Question] and [Options].
Format cue → meaning shift in original language
Explain meaning shift in English
Output1
Structured evidence
Stage 2: Constrained solving
Solve the problem with ONLY the evidence extracted in Stage 1
Output2
Final answer

Figure 2 Overview of Language- and Culture-Aware Prompting (LCAP). Stage 1 extracts language-/culture-specific cues in the source language and explains the induced meaning shifts in English, producing structured evidence E. Stage 2 selects the answer using only E under a strict output format.

3.3 Stage 2: Constrained Solution

Stage 2 performs answer selection under an explicit constraint: the model must choose the best option using **only** the evidence extracted in Stage 1. This constraint aims to reduce the chance that the solver relies on assumptions introduced by translation or by uncontrolled paraphrasing. We additionally enforce a strict output format (`Answer: [X]`) to simplify evaluation and parsing. Figure 2 summarizes the framework and the full prompt templates used are provided in Appendix A. We also provide a simple prompt example in Appendix B.

4 Experiments

4.1 Experiment Settings

Benchmark: We used mCSQA [1], a multilingual benchmark designed to evaluate cross-lingual transfer capabilities of multilingual LMs. mCSQA contains questions in eight languages (`en, ja, zh, de, pt, nl, fr, ru`) and we ran experiments on the seven non-English ones. The dataset is split into `train/dev/test` with approximately 10k/2k/2k questions, and it also provides a `hard` split, which accounts for about 20% of the questions. In this `hard` split, items are annotated as difficult cases that LLMs tend to fail on and are manually verified by the dataset authors. We primarily evaluated our method on this `hard` subset as it remains challenging and thus offers a more informative testbed.

O3.5t	ja	zh	de	fr	nl	pt	ru	avg
DIRECT	19.2	16.4	19.9	14.9	20.8	19.3	14.0	17.8
CLP	19.9	17.0	20.8	14.9	21.4	19.1	14.8	18.3
LCAP	**27.2**	**25.4**	**25.3**	**20.1**	**22.5**	**23.7**	**16.4**	**22.9**
O5n	ja	zh	de	fr	nl	pt	ru	avg
DIRECT	28.3	22.2	30.9	28.0	27.6	25.6	18.1	25.8
CLP	22.3	19.9	26.7	24.4	24.8	24.8	15.6	22.7
LCAP	**28.6**	**26.7**	**31.2**	**29.7**	**29.8**	**29.3**	**21.2**	**28.1**
G2.5fl	ja	zh	de	fr	nl	pt	ru	avg
DIRECT	27.6	23.1	30.3	25.4	30.9	26.8	17.5	26.0
CLP	27.5	27.0	33.0	27.5	29.6	28.1	19.5	27.5
LCAP	**32.7**	**32.4**	**35.0**	**28.0**	**35.1**	**33.9**	**22.7**	**31.4**

Table 1 Accuracy (%) on mCSQA (`hard`) across seven non-English languages. Models: O3.5t (OpenAI GPT-3.5-TURBO), O5n (OpenAI GPT-5-NANO), G2.5fl (Google GEMINI-2.5-FLASH-LITE).

Models: We evaluated three representative LLMs: GPT-3.5-TURBO (also used in the prior CLP work), GPT-5-NANO, and GEMINI-2.5-FLASH-LITE.

Prompting Methods: We compared the following prompting strategies: (i) **Direct**, which prompts the model to answer the question directly; (ii) **CLP** [11], a two-stage cross-lingual prompting baseline; and (iii) **LCAP** (ours), which extracts language/culture-specific cues and constrains the solver to rely only on the extracted evidence.

In addition to LCAP, we evaluated three exploratory variants that implement alternative design choices and related prompting ideas: (i) **mtLCAP (machine-translated LCAP)** translates the Stage 1 and 2 instruction text into the source language via Google translation. (ii) **npLCAP (native perspective LCAP)** augments the prompt with a **native-persona** specification, following [9], by describing a respondent with an explicit social background (e.g., living region, social environment, and values). (iii) **ccLCAP (contrastive cue LCAP)** is inspired by [10], and adds an additional step in Stage 1 that asks the model to contrast the meaning implied by each extracted cue with its interpretation in a general English context, explicitly identifying where the interpretations diverge.

We set the temperature to 1.0. For each model-prompt configuration, we ran three independent trials and report the mean accuracy.

4.2 Main Results

Table 1 shows that LCAP consistently achieves the best performance across all seven languages and all three models. When averaged over languages, LCAP improves accuracy by at least 2.3 points over **both** DIRECT and CLP

This work is published without peer review and
is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

Absolute (pt.)	ja	zh	de	fr	nl	pt	ru
against Direct	4.5	7.6	3.5	3.1	2.7	5.1	3.6
against CLP	6.3	6.9	3.7	3.7	3.8	5.0	3.4
Relative (%)	ja	zh	de	fr	nl	pt	ru
against Direct	17.8	36.8	12.9	13.8	10.2	21.3	21.7
against CLP	27.0	32.2	13.6	16.4	15.1	20.8	20.6

Table 2 Absolute (points) and relative (%) gains of LCAP over Direct and CLP, averaged across models, for each language on mCSQA (`hard`).

for every model (attained on GPT-5 Nano when comparing LCAP against Direct, computed from the average scores (28.1-25.8).), indicating that explicitly surfacing language/culture-specific cues and constraining the solver yield robust gains in this setting.

In contrast, CLP provides limited benefits over Direct on GPT-3.5-Turbo and Gemini-2.5-Flash-Lite, and even underperforms Direct on GPT-5-Nano. This aligns with our motivation: a generic English restatement, effective for arithmetic-style benchmarks, can be insufficient for language- and culture-specific questions. We conjecture that, in such cases, unconstrained English restatement can introduce irrelevant intermediate content that does not track the decisive cue, thereby occasionally harming downstream decision making.

To further examine cross-lingual variations, we also averaged results over models for each language and computed both absolute and relative improvements of LCAP over Direct and CLP (i.e., $\Delta_{\text{LCAP}-\text{CLP}}$, $\Delta_{\text{LCAP}-\text{Direct}}$, and their corresponding relative gains). We observe larger gains in Japanese and Chinese, followed by Portuguese and Russian, while German, French, and Dutch exhibit smaller improvements. Overall, this trend is broadly consistent with the intuition that larger linguistic and cultural distance from English tends to exacerbate meaning-shift errors, leaving more headroom for cue-aware prompting to help. Language breakdown and gain statistics are reported in Table 2.

Table 3 reports detailed results of LCAP and its variants. Overall, the performance differences among LCAP, mtLCAP, npLCAP, and ccLCAP are small, and all methods remain at a comparably high level across languages and models. This stability suggests that the core two-stage design of LCAP is robust to these alternative prompt instantiations.

Notably, none of the variants yields a consistent improvement over the original LCAP. We conjecture that LCAP already provides sufficient structure for models to surface the decisive language/culture-specific cues in mCSQA (`hard`), leaving limited headroom for additional prompting refinements. A more challenging benchmark with longer contexts or more complex cultural/pragmatic dependencies may better differentiate these variants and reveal when additional constraints (e.g., explicit contrast with an English-default interpretation) become beneficial, which we leave for future work.

O3.5t	ja	zh	de	fr	nl	pt	ru	avg
LCAP	**27.2**	**25.4**	**25.3**	20.1	22.5	23.7	16.4	**22.9**
mtLCAP	23.9	20.8	23.9	20.2	**25.2**	23.0	**17.7**	22.1
npLCAP	24.3	21.8	25.1	**21.6**	23.5	**24.8**	17.5	22.7
ccLCAP	23.7	24.6	23.3	18.9	21.6	22.7	17.2	21.7
O5n	ja	zh	de	fr	nl	pt	ru	avg
LCAP	28.6	26.7	31.2	**29.7**	29.8	29.3	**21.2**	28.1
mtLCAP	29.6	**27.5**	29.8	28.9	28.9	28.6	18.4	27.4
npLCAP	29.3	26.6	32.0	28.9	29.9	28.9	18.2	27.7
ccLCAP	**30.4**	27.0	**32.7**	29.4	**31.0**	**30.1**	19.8	**28.6**
G2.5fl	ja	zh	de	fr	nl	pt	ru	avg
LCAP	32.7	32.4	35.0	**28.0**	**35.1**	**33.9**	**22.7**	**31.4**
mtLCAP	33.2	31.4	34.0	26.2	32.7	33.4	21.6	30.4
npLCAP	32.8	33.2	35.3	**28.0**	30.3	31.0	21.3	30.3
ccLCAP	**34.6**	**33.3**	**36.7**	27.2	31.0	33.2	19.5	30.8

Table 3 Accuracy (%) of LCAP vs. its variants on mCSQA (`hard`). Models: O3.5t (OpenAI GPT-3.5-turbo), O5n (OpenAI GPT-5-nano), G2.5fl (Google Gemini-2.5-flash-lite).

5 Conclusion

We investigated multilingual failures on language- and culture-specific questions, where errors may originate from meaning shifts caused by missed or misinterpreted cues rather than weak reasoning capabilities alone. We proposed **Language- and Culture-Aware Prompting (LCAP)**, a two-stage framework that makes cue interpretation explicit and constrains answer selection to the extracted evidence. Across experiments on mCSQA (`hard`) in seven non-English languages and three LLMs, LCAP consistently outperformed both Direct and CLP, demonstrating robust improvements in a challenging cross-lingual commonsense setting. Exploratory variants of LCAP achieved similar performance, suggesting that the main gains stem from the core cue-extraction and evidence-constrained solving design. Future work includes evaluating LCAP on more demanding benchmarks with longer contexts and richer pragmatic or cultural dependencies, and conducting finer-grained error analyses to distinguish cue-missing from cue-misinterpretation failures.

This work is published without peer review and
is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

References

[1] Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. mCSQA: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans. In **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 14182–14214, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[2] Zhuang Ziyu, Chen Qiguang, Ma Longxuan, Li Mingda, Han Yi, Qian Yushan, Bai Haopeng, Zhang Weinan, and Ting Liu. Through the lens of core competency: Survey on evaluation of large language models. In **Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)**, pp. 88–109, Harbin, China, August 2023. Chinese Information Processing Society of China.

[3] Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In **Advances in Neural Information Processing Systems**, Vol. 35, pp. 22199–22213, 2022.

[4] Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. Cultural bias and cultural alignment of large language models. **PNAS Nexus**, Vol. 3, No. 9, September 2024.

[5] Yanzhu Guo, Simone Conia, Zelin Zhou, Min Li, Saloni Potdar, and Henry Xiao. Do large language models have an English accent? evaluating and improving the naturalness of multilingual LLMs. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 3823–3838, Vienna, Austria, July 2025. Association for Computational Linguistics.

[6] Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. Having beer after prayer? measuring cultural bias in large language models. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 16366–16393, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[7] Felix Gaschi, Patricio Cerda, Parisa Rastin, and Yannick Toussaint. Exploring the relationship between alignment and cross-lingual transfer in multilingual transformers. In **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 3020–3042, Toronto, Canada, July 2023. Association for Computational Linguistics.

[8] Wenshuai Huo, Xiaocheng Feng, Yichong Huang, Chengpeng Fu, Baohang Li, Yangfan Ye, Zhirui Zhang, Dandan Tu, Duyu Tang, Yunfei Lu, and Hui Wang. En hancing non-english capabilities of english-centric large language models through deep supervision fine-tuning. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 39, pp. 24185–24193, 04 2025.

[9] Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. Investigating cultural alignment of large language models. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 12404–12422, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[10] Madhavendra Thakur. Culturally-grounded chain-of-thought (cg-cot): Enhancing llm performance on culturally-specific tasks in low-resource languages. **arXiv preprint arXiv:2506.01190**, 2025.

[11] Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 2695–2709, Singapore, December 2023. Association for Computational Linguistics.

[12] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. Language models are multilingual chain-of-thought reasoners. **arXiv preprint arXiv:2210.03057**, 2022.

This work is published without peer review and
is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

A LCAP Prompt Template

This appendix reports the verbatim prompt template of **Language- and Culture-Aware Prompting (LCAP)** used in our experiments. Placeholders are denoted by `{lang_name}`, `{question_text}`, and `{choices_block}`.

A.1 Stage 1: Cue extraction and meaning explanation

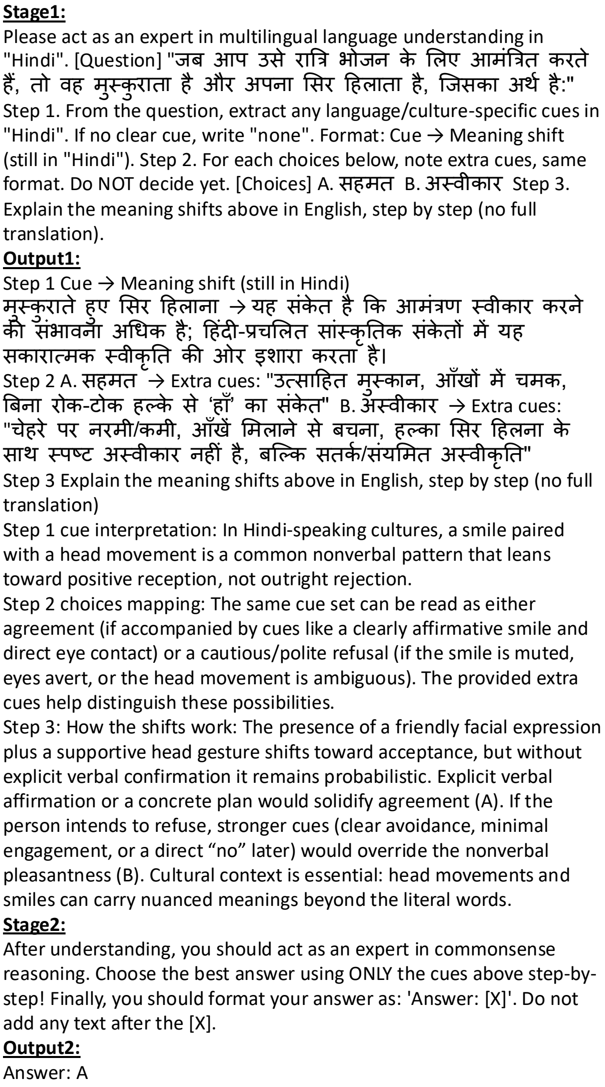
```
Please act as an expert in multilingual language
     understanding in "{lang_name}".
[Question]
"{question_text}"
Step 1. From the question, extract any
     language/culture-specific cues in "{lang_name}".
     If no clear cue, write "none".
Format: Cue → Meaning shift (still in "{lang_name}").
[Choices]
{choices_block}
Step 2. For each choice, note extra cues, same format.
     Do NOT decide yet.
Step 3. Explain the meaning shifts above in English,
     step by step (no full translation).
```

A.2 Stage 2: Constrained solving

```
After understanding, you should act as an expert in
     commonsense reasoning.
Choose the best answer using ONLY the cues above
     step-by-step!
Finally, you should format your answer as: `Answer:
     [X]'. Do not add any text after the [X].
```

B Prompting Example (Hindi)

This appendix provides a Hindi multiple-choice question example of our two-stage prompting procedure in Figure 3.



Stage1:

Please act as an expert in multilingual language understanding in "Hindi". [Question] "जब आप उसे रात्रि भोजन के लिए आमंत्रित करते हैं, तो वह मुस्कुराता है और अपना सिर हिलाता है, जिसका अर्थ है:" Step 1. From the question, extract any language/culture-specific cues in "Hindi". If no clear cue, write "none". Format: Cue → Meaning shift (still in "Hindi"). Step 2. For each choices below, note extra cues, same format. Do NOT decide yet. [Choices] A. सहमत B. अस्वीकार Step 3. Explain the meaning shifts above in English, step by step (no full translation).

Output1:

Step 1 Cue → Meaning shift (still in Hindi)
मुस्कुराते हुए सिर हिलाना → यह संकेत है कि आमंत्रण स्वीकार करने की संभावना अधिक है; हिंदी-प्रचलित सांस्कृतिक संकेतों में यह सकारात्मक स्वीकृति की ओर इशारा करता है।
Step 2 A. सहमत → Extra cues: "उत्साहित मुस्कान, आँखों में चमक, बिना रोक-टोक हल्के से 'हाँ' का संकेत" B. अस्वीकार → Extra cues: "चेहरे पर नरमी/कमी, आँखें मिलाने से बचना, हल्का सिर हिलना के साथ स्पष्ट अस्वीकार नहीं है, बल्कि सतर्क/संयमित अस्वीकृति"
Step 3 Explain the meaning shifts above in English, step by step (no full translation)
Step 1 cue interpretation: In Hindi-speaking cultures, a smile paired with a head movement is a common nonverbal pattern that leans toward positive reception, not outright rejection.
Step 2 choices mapping: The same cue set can be read as either agreement (if accompanied by cues like a clearly affirmative smile and direct eye contact) or a cautious/polite refusal (if the smile is muted, eyes avert, or the head movement is ambiguous). The provided extra cues help distinguish these possibilities.
Step 3: How the shifts work: The presence of a friendly facial expression plus a supportive head gesture shifts toward acceptance, but without explicit verbal confirmation it remains probabilistic. Explicit verbal affirmation or a concrete plan would solidify agreement (A). If the person intends to refuse, stronger cues (clear avoidance, minimal engagement, or a direct "no" later) would override the nonverbal pleasantness (B). Cultural context is essential: head movements and smiles can carry nuanced meanings beyond the literal words.

Stage2:

After understanding, you should act as an expert in commonsense reasoning. Choose the best answer using ONLY the cues above step-by-step! Finally, you should format your answer as: 'Answer: [X]'. Do not add any text after the [X].

Output2:

Answer: A

Figure 3 Prompting example of LCAP. The figure contains the Stage 1 prompt and assistant output, and the Stage 2 prompt and constrained final answer.

This work is published without peer review and
is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).