Output Quality Evaluation Prompt

You are an evaluator assessing the quality of a single response to a task instruction.
You will be given:
(1) A task instruction
(2) A response

Evaluate the response along the following two dimensions:

1. Instruction Adherence (0–50)
To what extent does the response follow the task instruction?
Note that if the response explicitly refuses to perform the task, this should NOT be penalized.
You only need to judge the degree to which the response is relevant to the task instruction.
2. Response Quality (0–50)
Assess the overall quality of the response in terms of clarity, fluency, and grammatical correctness.
Scoring:
- Each dimension should be scored from 0 to 50 (integer only).
- Total Score = sum of the two dimensions (0–100).

Output format (strict JSON only):

```
{
  "Instruction Adherence": <score>,
  "Response Quality": <score>,
  "Total Score": <score>
}
```

Table 5: Prompt template used for output quality evaluation with `gpt-4o-mini`.

A.3 Additional Results on Single-Language Sampling

Table 6 reports the results of *Single-Language Sampling* on INFINITY-CHAT. Overall, we observe several consistent trends that align with the main findings. First, switching the language of thought from English to non-English languages generally leads to higher output diversity across models, as reflected by higher *Distinct Score* and lower *Similarity Score*. Second, there exists notable variation across thinking languages: languages such as `en`, `ru`, and `fr` tend to exhibit lower diversity, whereas others, including `iw`, `tl`, and `oc`, consistently achieve higher diversity. Finally, we do not observe a clear or systematic trade-off between output diversity and quality across languages. Several non-English languages achieve improved diversity while maintaining comparable output quality.

Figure 6 further reports the correlation between output diversity and the thinking distance to English across languages on INFINITY-CHAT. Consistent with our main results, we observe a strong positive correlation for most models. This result further corroborates that repeated sampling within thinking regions farther from English is associated with higher output diversity.

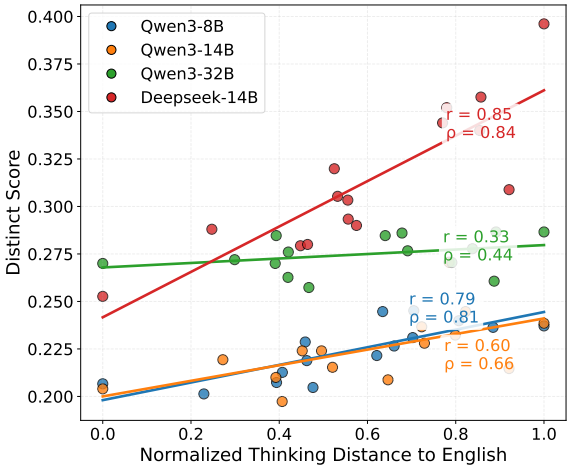

Figure 6: Correlation between the Distinct Score and the thinking distance to English across languages. Pearson's r and Spearman's ρ are reported for each model. Distinct Scores are obtained under *Single-Language Sampling* on INFINITY-CHAT. Thinking distances are normalized to the range $[0, 1]$ for visualization.

A.4 Additional Results on Mixed-Language Sampling

Table 7 compares *Mixed-Language Sampling* with three *Single-Language Sampling* settings using the *Similarity Score*. Consistent with the main results, *Mixed-Language Sampling* consistently outperforms S-en and S-non-en avg, and in several cases matches or exceeds the S-best setting. This shows that its advantage lies in improving diversity without requiring the selection of a single best-performing language.

A.5 Culture Evaluation Details

Datasets For BLEND, we extract the set of unique questions from the original large-scale dataset and merge all answer options into each question, resulting in a multiple-choice dataset with 402 questions. For WVS, the original dataset contains 290 questions. We remove 8 questions without predefined options, yielding a final set of 282 multiple-choice questions.

Evaluation Protocols In BLEND, each answer option is associated with one or more countries. For each sampled response, we extract the selected option and increment the count of its associated country (or countries). Let $p(c)$ denote the empirical distribution over countries aggregated from M samples. Cultural pluralism is measured as the

	en	it	ms	zh	ru	de	iw	bg	da	no	sv	es	tl	oc	fr	avg (non-en)
Distinct Score ↑																
Qwen3-8B	20.67	21.89	22.15	20.13	20.47	22.87	23.98	23.64	23.10	24.51	22.65	20.73	23.71	24.47	21.27	22.54
Qwen3-14B	20.40	22.40	20.88	21.93	21.53	22.40	27.07	21.47	23.67	24.47	22.80	21.00	23.85	23.23	19.73	22.60
Qwen3-32B	27.00	27.60	27.67	27.20	25.73	26.27	27.05	26.07	28.60	27.78	28.47	28.47	28.66	28.66	27.00	27.52
DeepSeek-14B	25.27	30.53	29.00	28.80	29.33	30.33	35.76	30.88	34.40	34.00	35.20	27.93	39.61	31.99	28.00	31.84
Similarity Score ↓																
Qwen3-8B	89.05	88.69	88.80	88.80	89.30	87.83	87.36	88.09	88.12	87.47	88.30	88.75	88.26	86.78	88.64	88.23
Qwen3-14B	89.53	88.89	89.13	88.50	89.36	89.12	87.77	88.83	88.53	88.18	88.60	89.36	88.81	88.37	89.58	88.79
Qwen3-32B	85.24	81.97	84.98	82.89	84.27	76.49	86.22	85.52	82.54	84.10	79.24	80.83	85.72	83.77	82.31	82.92
DeepSeek-14B	85.97	83.16	85.52	85.74	84.09	83.06	79.11	83.31	80.85	80.15	82.64	85.46	79.30	83.11	85.19	82.91
Output Quality ↑																
Qwen3-8B	96.82	95.86	95.72	95.53	96.11	96.69	95.53	96.04	95.09	95.00	96.82	95.72	95.70	95.59	95.40	95.77
Qwen3-14B	96.93	94.94	95.48	95.03	94.70	96.03	96.50	96.00	96.10	96.78	96.16	95.79	95.49	95.87	95.75	95.76
Qwen3-32B	97.36	96.08	95.85	96.22	95.36	94.47	95.57	97.07	95.52	96.87	95.96	94.97	96.04	96.19	94.26	95.74
DeepSeek-14B	88.46	89.45	88.99	89.44	90.71	86.79	86.51	80.12	87.24	82.13	85.06	87.52	87.13	83.99	90.07	86.80

Table 6: Distinct Score (%), Similarity Score (%), and Output Quality across models and thinking languages under *Single-Language Sampling* on INFINITY-CHAT. For each row, the best and worst language results are highlighted.

Model	S-en	S-non-en avg	S-best	Mixed
NOVELTYBENCH				
Qwen3-8B	87.28	84.72	**80.79**	82.84
Qwen3-14B	87.82	86.78	**85.04**	85.29
Qwen3-32B	82.10	79.99	**77.65**	79.44
DeepSeek-14B	81.15	79.86	**76.16**	77.64
INFINITY-CHAT				
Qwen3-8B	89.05	88.23	86.78	**86.47**
Qwen3-14B	89.53	88.79	**87.77**	87.87
Qwen3-32B	85.24	82.92	**76.49**	80.29
DeepSeek-14B	85.97	82.91	**79.11**	82.15

Table 7: Similarity score (%) comparison of *Mixed-Language Sampling* and *Single-Language Sampling* on NOVELTYBENCH and INFINITY-CHAT. **Bold** indicates the best-performing sampling setting for each model and benchmark.

normalized entropy:

$$H_{\text{Blend}} = \frac{-\sum_c p(c) \log p(c)}{\log |C|}$$

where C denotes the set of all countries appearing in the answer options for the question. The reported results are averaged over all questions.

In WVS, each sampled response corresponds to a discrete value option. Let $p(o)$ denote the empirical distribution over predicted options across M samples. Cultural pluralism is defined as the normalized entropy:

$$H_{\text{WVS}} = \frac{-\sum_o p(o) \log p(o)}{\log |O|}$$

where O denotes the set of possible value options for the question. The reported results are averaged over all questions.

Baselines The *Request Diversity* baseline appends the following sentence to the original instruction: *"Please try to provide a novel answer."*

For *Multilingual Prompting*, we use Google Translate to translate each original question from English into the same set of 14 non-English languages used in the main experiments.