Query-Following vs Context-Anchoring: How LLMs Handle Cross-Turn Language Switching

Kyuhee Kim, Chengheng Li Chen, Anna Sotnikova
EPFL, Lausanne, Switzerland
kyuhee.kim@epfl.ch, chengheng.lichen@epfl.ch, anna.sotnikova@epfl.ch

Abstract

When multilingual users switch languages mid-conversation, how should LLMs respond? We extend MultiChallenge to evaluate cross-turn language switching, translating 182 multi-turn conversations into German, Chinese, Spanish, and Arabic. Across five frontier models, we observe asymmetric behavior: switching into a foreign language (EN→X) yields high query-language fidelity (89–99%), but switching back to English (X→EN) reveals divergent policies. GPT-5 follows the query language (>95%), while Claude Opus 4.5 and Command R+ maintain the established conversation language (<8%). Task accuracy remains stable across conditions regardless of language selection differences. A simple explicit system prompt shows limited effectiveness in modifying these defaults. Our code and data are available at https://github.com/koreankiwi99/crossturn-lang-switch.

1 Introduction

Multilingual speakers frequently switch languages within conversations, beginning a query in English, then continuing in their native language, or vice versa. Over half the world's population is multilingual, and this cross-turn language switching reflects natural communication patterns. Yet how LLMs respond when users switch languages mid-conversation remains unexplored. A model could follow the query language, respecting the user's immediate choice, or anchor to the established conversation language, maintaining consistency.

Prior work examines multilingual and multi-turn capabilities separately. Language confusion research focuses on single-turn interactions (Marchisio et al., 2025), multi-turn benchmarks operate monolingually (Zheng et al., 2023; Sirdeshmukh et al., 2025), and multilingual benchmarks keep each conversation in one language (He et al., 2024). Cross-turn language switching, where users change languages mid-conversation, remains untested.

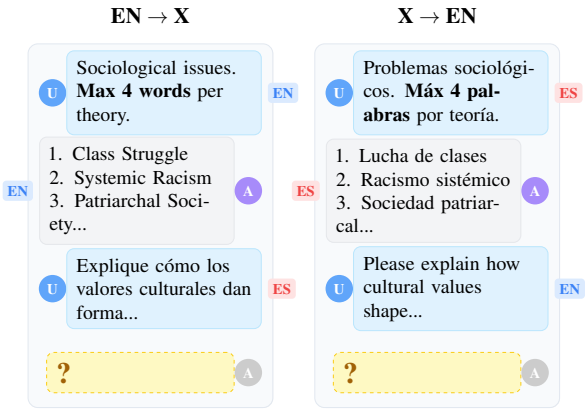

Figure 1: Same conversation in opposite switching directions. **EN→X**: English context with Spanish query. **X→EN**: Spanish context with English query. Both require the model to maintain the 4-word constraint while switching languages.

We address this gap by extending MultiChallenge to evaluate cross-turn language switching. We translate the INFERENCE_MEMORY and INSTRUCTION_RETENTION categories, which require cross-turn information retention, into German, Chinese, Spanish, and Arabic. We test four conditions: monolingual baselines in English and X, plus two switching conditions where the language changes at the user's final turn (Figure 1). We evaluate five frontier models and measure two dimensions: **Language Fidelity** (does the model respond in the user's language?) and **Task Accuracy** (does switching degrade performance?).

Our results reveal asymmetric behavior. When switching into a foreign language (EN → X), all models follow the query language (89–99%). When switching back to English (X → EN), models diverge: GPT-5 follows the query (>95%), while Claude Opus 4.5 and Command R+ continue in the foreign language (<8%). Task accuracy remains stable even as fidelity diverges, suggesting models comprehend cross-lingual input but apply different language policies. A simple system prompt

explicitly instructing query-language response fails to override these defaults, suggesting this behavior may not be easily modifiable at inference time.

These findings highlight the need for systematic research on user preferences for language continuity, standardized evaluation of cross-lingual behavior, and configurable language policies in multilingual deployment. We contribute: (1) the first systematic evaluation of cross-turn language switching in multi-turn conversations (182 conversations $\times$ 4 languages), (2) evidence that frontier models implement fundamentally different language policies, a divergence invisible to existing evaluations, and (3) demonstration that task accuracy remains stable despite divergent language behavior, isolating this as a policy rather than capability issue.

2 Related Work

Prior work has examined multilingual capabilities, multi-turn interactions, and code-switching separately. Yet no benchmark tests how models handle language switches across conversation turns.

Language Confusion. Marchisio et al. (2025) introduced the Language Confusion Benchmark, which measures whether models respond in the language they were prompted in. They found that Llama and Mistral models frequently respond in unintended languages, especially in cross-lingual settings. Nie et al. (2025) traced this to transition failures in final layers; Lee et al. (2025) mitigated it via preference tuning. However, all prior work focuses on single-turn interactions.

Multi-Turn and Multilingual Benchmarks. MT-Bench (Zheng et al., 2023) introduced multi-turn evaluation with open-ended conversational questions. MultiChallenge (Sirdeshmukh et al., 2025) tests more demanding scenarios such as retaining first-turn instructions or recalling user information across turns. Both operate monolingually. Multi-IF (He et al., 2024) extends instruction-following evaluation to 8 languages with 3-turn conversations, but each conversation remains in a single language throughout.

Code-Switching and User Expectations. Research on code-switching in NLP has primarily focused on intra-sentential mixing, where speakers blend languages within a single utterance. Benchmarks like LinCE (Aguilar et al., 2020) and GLUE-CoS (Khanuja et al., 2020) evaluate model performance on such naturally code-switched text. However, Human-Computer Interaction (HCI) research shows that multilingual users also switch languages across conversation turns and expect agents to accommodate this. Choi et al. (2023) found that code-mixing users feel excluded when conversational agents cannot handle their language practices. Bawa et al. (2020) showed that bilingual users prefer chatbots that reciprocate their language choices over multiple turns. Despite these documented user behaviors, no work evaluates how LLMs respond to cross-turn language switching.

3 Methodology

We design a controlled framework to evaluate language model behavior under mid-conversation language switches. Our approach comprises dataset construction through benchmark translation, experimental conditions isolating switch effects, and dual metrics assessing both language fidelity and task accuracy.

3.1 Dataset Construction

We extend MultiChallenge's INFERENCE_MEMORY (113 examples), which tests whether models recall and connect user information scattered across previous turns, and INSTRUCTION_RETENTION (69 examples), which tests whether models follow first-turn instructions throughout the conversation. Conversations range from 3 to 19 turns (median: 7): short (3–5 turns, n=44), medium (7–9 turns, n=99), and long (11+ turns, n=39).

We translate conversations into German, Chinese, Spanish, and Arabic using Google Translate, with human verification on sampled translations and GPT-4o-mini automated verification across all translations (Appendix A).

3.2 Experimental Conditions

Each MultiChallenge conversation contains multiple context turns followed by a final evaluation query. We test four conditions (Table 1), including monolingual baselines in English and X, plus two switching conditions where the language changes at the user's final turn.

3.3 Models

We evaluate five frontier models: GPT-5 (OpenAI, 2025), Gemini 3 Pro (Google, 2025), Claude Opus 4.5 (Anthropic, 2025), DeepSeek-V3.1 (DeepSeek-AI, 2025), and Command R+ (Cohere, 2024). All models are accessed via API in non-thinking mode.

Condition	Description
Baseline (EN)	All turns in English (original)
Baseline (X)	All turns in X (translated)
EN→X	English context, final query in X
X→EN	X context, final query in English

Table 1: Experimental conditions where X ∈ {DE, ZH, ES, AR}. Context refers to all user–assistant turns preceding the final query.

Specific model versions and parameters are detailed in Appendix B.

3.4 Evaluation Metrics

We measure two dimensions: (1) **Language Fidelity**, the percentage of responses matching the query language, evaluated using GPT-4o-mini as judge; and (2) **Task Accuracy**, whether the response correctly addresses the task, evaluated using GPT-4o with MultiChallenge's instance-level rubrics. Details of the evaluation prompts are provided in Appendix C.

4 Results

Following the methodology outlined above, we present our findings on language fidelity and task accuracy. Our results reveal some differences in how models handle mid-conversation language switches, suggesting distinct underlying response strategies.

4.1 Language Fidelity

Table 2 presents our main finding: models diverge dramatically in the X→EN condition (foreign context, English query).

Model	EN→X	X→EN	Behavior
GPT-5	98.6	**95.1**	Query-following
Gemini 3 Pro	98.3	73.8	Mixed
Claude Opus 4.5	96.1	7.7	Context-anchoring
DeepSeek-V3.1	88.3	51.9	Mixed
Command R+	89.3	**0.8**	Context-anchoring

Table 2: Language fidelity (%) by condition, averaged across languages. Bold indicates extreme values.

All models successfully follow the query language when switching into a foreign language (EN→X: 88–99%). However, when switching back to English after foreign context (X→EN), models split into three groups:

Query-following GPT-5 responds in the query language regardless of context (95.1% English).

Context-anchoring Claude Opus 4.5 and Command R+ continue in the context language, largely ignoring the language switch (0.8–7.7% English).

Mixed Gemini 3 Pro and DeepSeek-V3.1 show intermediate behavior, balancing query and context influence (51.9–73.8%).

These reflect different design choices rather than performance differences.

4.2 Per-Language Analysis

Table 3 breaks down X→EN fidelity by source language.

Model	DE	ZH	ES	AR	Avg
GPT-5	94.0	95.6	94.5	96.2	95.1
Gemini 3 Pro	78.6	72.5	74.7	69.2	73.8
Claude Opus 4.5	10.4	9.9	6.0	4.4	7.7
DeepSeek-V3.1	41.8	60.4	41.2	64.3	51.9
Command R+	1.1	1.1	0.5	0.5	0.8

Table 3: X→EN fidelity (%) by source language.

GPT-5 maintains consistent fidelity across languages (94–96%). DeepSeek-V3.1 shows notable variation: higher fidelity when switching from Chinese (60.4%) and Arabic (64.3%) compared to German (41.8%) and Spanish (41.2%), possibly reflecting training data composition. Context-anchoring models (Claude, Command R+) show uniformly low fidelity regardless of source language.

4.3 Conversation Length Effect

Table 4 analyzes whether context-anchoring intensifies with conversation length. We use chi-square tests to assess whether fidelity rates differ significantly across length categories.

Model	Short	Med	Long	p
GPT-5	97.2	93.9	95.5	0.25
Gemini 3 Pro	82.4	75.0	60.9	$<$0.001
Claude Opus 4.5	11.9	5.8	7.7	0.04
DeepSeek-V3.1	55.1	54.0	42.9	0.04
Command R+	0.0	0.5	2.6	0.02

Table 4: X→EN fidelity (%) by conversation length. Short: 3–5 turns (n=44), Medium: 7–9 turns (n=99), Long: 11+ turns (n=39). p-values from χ^2 tests.

Gemini 3 Pro shows significant degradation with length, dropping from 82.4% to 60.9% (p $<$0.001). GPT-5 remains stable across all lengths (p=0.25, not significant). Context-anchoring models show floor effects, with fidelity already near zero, leaving no room for further decline.

4.4 Task Accuracy

Table 5 shows task accuracy remains stable across conditions, with no significant degradation from language switching.

Model	Base EN	Base X	EN→X	X→EN
GPT-5	57.1	58.7	59.2	52.5
Gemini 3 Pro	71.4	70.1	70.9	70.4
Claude Opus 4.5	54.4	48.5	48.9	49.9
DeepSeek-V3.1	50.0	40.3	42.9	37.7
Command R+	15.9	11.3	15.0	11.4

Table 5: Task accuracy (%) by condition, averaged across languages.

This null result is informative as language switching does not impair task performance. The challenge is behavioral (which language to use), not comprehension (understanding the task). Full results are provided in Appendix D.

5 Analysis

To further understand model behavior, we conduct additional analyses examining non-English language pairs and the influence of explicit system prompt instructions on language fidelity.

5.1 Non-English Switching (X→Y)

Table 6 presents results for switching between non-English languages, examining whether English has a privileged role. We select four pairs representing diverse script combinations.

Model	Metric	ZH→DE	DE→ZH	ES→AR	AR→ES
GPT-5	Fidelity	96.2	97.3	96.2	98.4
	Accuracy	53.8	52.2	56.6	51.6
Claude Opus 4.5†	Fidelity	64.3	35.2	81.3	19.2
	Accuracy	47.3	46.2	51.1	48.9

Table 6: Cross-lingual transfer (X→Y) results (%). †One empty response in ZH→DE (n=181).

GPT-5 maintains high fidelity (>96%) across all X→Y pairs, consistent with its query-following behavior. Claude Opus 4.5 shows variable fidelity (19–81%), suggesting language-specific biases beyond simple context-anchoring. Task accuracy remains stable for both models, reinforcing that language switching does not impair performance.

5.2 System Prompt Ablation

Table 7 presents results testing whether explicit instructions ("Always respond in the language of the user's most recent message") can override default behavior.

Model	Condition	None	Explicit
GPT-5	X→EN	95.1	94.0
Claude 4.5	X→EN	7.7	7.1
Command R+	X→EN	0.8	0.7

Table 7: X→EN fidelity (%) with and without explicit language instructions. Full results in Appendix E.

Explicit system prompts have minimal effect (<1.5 pp difference), suggesting language behavior is deeply embedded and not easily overridden by simple instructions.

6 Discussion

Implications for Multilingual UX. Our findings reveal a design tension. Context-anchoring may frustrate users who intentionally switch languages, while query-following may feel inconsistent to users who prefer conversational continuity. The appropriate behavior depends on user intent, whether the switch was deliberate or incidental, which current models cannot infer.

Why the Divergence? We hypothesize the divergence stems from different training objectives. Context-anchoring may result from training models to maintain consistency and avoid sycophancy (following user requests unconditionally). Query-following may prioritize immediate user intent. Our system prompt ablation (§5.2) shows that explicit instructions fail to override these defaults, though the underlying mechanism remains unclear. Future work using mechanistic interpretability or reasoning trace analysis could help identify where and how language selection occurs during generation.

Future Directions. Several questions remain open. First, at what point does accumulated foreign context override query-language signals? Varying switch points within conversations may reveal thresholds for context-anchoring. Second, tasks involving culture-specific knowledge may show accuracy degradation unlike our general reasoning tasks. Third, user preference studies could determine whether multilingual users prefer query-following or context-anchoring behavior in practice.

7 Conclusion

We evaluate cross-turn language switching in multi-turn conversations, revealing that frontier models

diverge in language policy: GPT-5 follows user language switches (>95%), while Claude Opus 4.5 and Command R+ anchor to conversation context (<8%). Task accuracy remains stable across conditions, indicating the challenge is behavioral rather than cognitive. These findings have implications for multilingual deployment and model selection.

Limitations

Our study has several limitations. First, we use pre-filled conversation histories following standard practice (Sirdeshmukh et al., 2025; Zheng et al., 2023), which may not fully reflect live interaction dynamics. Second, translation quality, while verified, may introduce artifacts. Third, we test four high-resource languages; low-resource languages may reveal different patterns. Fourth, task accuracy shows moderate cross-run variance (std up to 3 pp; Tables 11 and 12), so absolute values should be interpreted cautiously. Fifth, comprehensive consistency analysis is limited to two models (GPT-5, Gemini 3 Pro) on one condition (ES→EN) due to computational constraints.

Ethics Statement

This work uses publicly available benchmark data (MultiChallenge) and commercially available language models accessed through their official APIs. Human verification of translations was conducted by native speakers who participated voluntarily. The research does not involve sensitive or private data, and we do not foresee direct negative societal impacts from this work.

References

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France. European Language Resources Association.

Anthropic. 2025. Introducing claude opus 4.5. https://www.anthropic.com/news/claude-opus-4-5.

Anshul Bawa, Pranav Khadpe, Pratik Joshi, Kalika Bali, and Monojit Choudhury. 2020. Do multilingual users prefer chat-bots that code-mix? let's nudge and find out! *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1).

{Yunjae J.} Choi, Minha Lee, and Sangsu Lee. 2023. Toward a multilingual conversational agent: Challenges and expectations of code-mixing multilingual users. In *CHI 2023 - Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Conference on Human Factors in Computing Systems - Proceedings. Association for Computing Machinery. Publisher Copyright: © 2023 ACM.; 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023 ; Conference date: 23-04-2023 Through 28-04-2023.

Cohere. 2024. Command r+. https://docs.cohere.com/docs/command-r-plus.

DeepSeek-AI. 2025. Deepseek-v3.1. https://huggingface.co/deepseek-ai/DeepSeek-V3.1.

Google. 2025. Gemini 3: Introducing the latest gemini ai model from google. https://blog.google/products/gemini/gemini-3/.

Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, Shruti Bhosale, Chenguang Zhu, Karthik Abinav Sankararaman, Eryk Helenowski, Melanie Kambadur, Aditya Tayade, Hao Ma, Han Fang, and Sinong Wang. 2024. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following. *Preprint*, arXiv:2410.15553.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online. Association for Computational Linguistics.

Nahyun Lee, Yeongseo Woo, Hyunwoo Ko, and Guijin Son. 2025. Controlling language confusion in multilingual LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1026–1035, Vienna, Austria. Association for Computational Linguistics.

Kelly Marchisio, Wei-Yin Ko, Alexandre Bérard, Théo Dehaze, and Sebastian Ruder. 2025. Understanding and mitigating language confusion in llms. *Preprint*, arXiv:2406.20052.

Ercong Nie, Helmut Schmid, and Hinrich Schütze. 2025. Mechanistic understanding and mitigation of language confusion in english-centric large language models. *Preprint*, arXiv:2505.16538.

OpenAI. 2025. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/.

Ved Sirdeshmukh, Kaustubh Deshpande, Johannes Mols, Lifeng Jin, Ed-Yeremai Cardona, Dean Lee, Jeremy Kritz, Willow Primack, Summer Yue, and Chen Xing. 2025. Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms. *Preprint*, arXiv:2501.17399.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Translation and Verification Details

A.1 Translation Pipeline

We translate all 182 conversations from English into German (DE), Chinese (ZH), Spanish (ES), and Arabic (AR) using Google Translate. Each conversation was translated in full, preserving turn structure and speaker labels.

A.2 Human Verification

To assess translation quality, native speakers verified 10 randomly sampled translations each for Spanish and Chinese (Table 8), scoring on semantic accuracy, completeness, and overall quality (1–5 scale).

Common issues identified in Spanish included literal translations of idiomatic expressions (e.g., "grounding techniques" → "técnicas de conexión a tierra" instead of the more natural "técnicas de anclaje"). For Chinese, issues included mistranslation of domain-specific terms and inconsistent handling of proper nouns. These errors informed our automated verification prompt but were not individually corrected in the final dataset. Despite these minor issues, both languages achieved perfect completeness scores, indicating no information loss.

Language	Semantic	Complete	Overall	n
Spanish	4.1	5.0	4.4	10
Chinese	4.9	5.0	4.8	10

Table 8: Human verification scores (1–5 scale).

A.3 Automated Verification and Correction

Based on error patterns from human verification, we develop a GPT-4o-mini prompt to verify and correct task-critical elements across all translations. German and Arabic translations were verified and corrected using this automated approach only. Table 9 summarizes results by language.

All experiments use the corrected translations from both human and automated verification.

Language	Turns	Corrected	Rate
German	1,460	265	18.2%
Chinese	1,460	254	17.4%
Spanish	1,460	257	17.6%
Arabic	1,460	271	18.6%
Total	5,840	1,047	17.9%

Table 9: GPT-4o-mini automated verification results (turn-level).

A.4 Verification Prompt

System Prompt

You are a translation quality assessor. Compare an original English conversation with its translation and assess whether the translation accurately conveys the original meaning. CRITICAL: The translation must match exactly what the original says, word-for-word. Do NOT correct inconsistencies or errors in the original text — translate them literally as written.

User Prompt Template

```
Compare the original English conversation
with its {target_language} translation.
Assess whether each turn accurately conveys
the original meaning. Focus on:
- Entity accuracy (names, places,
activities, objects)
- Numerical accuracy (dates, times,
quantities)
- Constraint compliance (formatting
requirements)
- Semantic fidelity (meaning preservation)
Output JSON:
{
  "accurate": <true|false>,
  "issues": [{"turn": <n>, "description":
"<what is wrong>",
    "original": "<text>", "translated":
"<text>", "corrected": "<text>"}],
  "corrected_conversation": [{"role":
"user", "content": "<content>"}, ...]
}
The corrected_conversation must have
exactly {turn_count} turns.
If accurate, corrected_conversation should
be a copy of the translated conversation.
ORIGINAL (English):
{original_conversation}
TRANSLATED ({target_language}):
{translated_conversation}
```

B Model Settings

Table 10 lists the specific model versions and API parameters used in our experiments.

GPT-5 and Gemini 3 Pro do not support custom temperature values; we use their defaults (temperature = 1). For other models, we set temperature to 0. Variance checks on ES→EN (n=182, 3

Model	Version	Temperature
GPT-5	gpt-5-2025-08-07	1 (default)
Gemini 3 Pro	gemini-3.0-pro-preview-2025-11-18	1 (default)
Claude Opus 4.5	claude-opus-4-5-20251101	0
DeepSeek-V3.1	deepseek-v3-1-250821	0
Command R+	command-r-plus-08-2024	0

Table 10: Model versions and API settings. GPT-5 and Gemini 3 Pro do not support custom temperature values for reasoning tasks.

runs) for GPT-5 and Gemini 3 Pro are shown in Tables 11 and 12. Language fidelity is stable (std <1 pp), while task accuracy shows moderate variance (std up to 3 pp), reflecting model stochasticity and LLM-based evaluation noise. Comprehensive consistency analysis across all models and conditions remains future work.

Model	Mean (%)	Std	κ	3-Way (%)
GPT-5	94.9	0.52	0.74	96.2
Gemini 3 Pro	74.5	0.26	0.64	79.7

Table 11: Language fidelity consistency (ES→EN, 3 runs). κ = Cohen's kappa (mean pairwise).

Model	Mean (%)	Std	κ	3-Way (%)
GPT-5	50.2	0.52	0.55	66.5
Gemini 3 Pro	71.1	2.99	0.64	77.5

Table 12: Task accuracy consistency (ES→EN, 3 runs). κ = Cohen's kappa (mean pairwise).

C Evaluation Details

C.1 Language Fidelity

We use GPT-4o-mini (`gpt-4o-mini`) as a language identification judge. Given a model response and the expected target language, the judge determines whether the response is written in the expected language.

Language Fidelity Prompt

Is this text written in {expected_lang_name}?
<TEXT>
{response}
</TEXT>
Rules:
- Words like "No", "OK" exist in multiple languages — consider them valid in the expected language
- Ignore code snippets, URLs, or proper nouns
- Focus on the main content language
Return ONLY: YES or NO
Your answer:

C.2 Task Accuracy

We use GPT-4o (`gpt-4o-2024-08-06`) as judge following MultiChallenge's LLM-as-judge protocol with instance-level rubrics. Each test instance includes a task-specific rubric that defines correctness criteria. The judge receives the conversation context, model response, and rubric, then outputs a binary correctness judgment. This approach achieves 93% alignment with human raters.

Task Accuracy Prompt

You are tasked with evaluating a model response to see if it meets a specific criteria. The criteria will always be YES/NO evaluation.
The model response is as follows:
<MODEL_RESPONSE>
{response}
</MODEL_RESPONSE>
The criteria that the model response must meet is as follows. Be VERY STRICT!:
<CRITERIA>
{target_question}
</CRITERIA>
Print your reasoning followed by your verdict, either "YES" or "NO".

D Full Results

Tables 13 and 14 report language fidelity and task accuracy across all conditions and languages.Empty responses were excluded: Gemini 3 Pro (8), Claude Opus 4.5 (3).

E System Prompt Ablation

Tables 15 and 16 report full results for the system prompt ablation. Neither language fidelity nor task accuracy changes meaningfully with explicit instructions.

	Base EN	Baseline X				EN→X				X→EN			
Model		DE	ZH	ES	AR	DE	ZH	ES	AR	DE	ZH	ES	AR
GPT-5	100.0	98.9	100.0	100.0	99.5	97.8	99.5	99.5	97.8	94.0	95.6	94.5	96.2
Gemini 3 Pro	100.0	98.9	100.0	100.0	99.5	98.3	98.9	98.4	97.8	78.6	72.5	74.7	69.2
Claude Opus 4.5	100.0	98.9	100.0	100.0	99.5	96.7	94.0	97.3	96.7	10.4	9.9	6.0	4.4
DeepSeek-V3.1	100.0	98.9	98.4	100.0	98.9	93.4	73.1	95.1	91.8	41.8	60.4	41.2	64.3
Command R+	100.0	98.9	100.0	100.0	99.5	91.8	89.0	95.6	80.8	1.1	1.1	0.5	0.5

Table 13: Language fidelity (%).

	Base EN	Baseline X				EN→X				X→EN			
Model		DE	ZH	ES	AR	DE	ZH	ES	AR	DE	ZH	ES	AR
GPT-5	57.1	58.2	57.7	57.7	61.0	57.1	59.9	59.3	60.4	55.5	50.5	49.5	54.4
Gemini 3 Pro	71.4	66.5	72.0	71.4	70.3	73.6	70.3	68.7	70.9	66.5	68.7	72.0	74.2
Claude Opus 4.5	54.4	45.1	48.9	52.7	47.3	49.5	46.7	50.5	48.9	48.4	47.8	52.7	50.5
DeepSeek-V3.1	50.0	39.0	39.0	45.1	37.9	40.1	44.5	44.0	42.9	38.5	37.4	37.9	36.8
Command R+	15.9	11.5	9.3	9.9	14.3	15.4	13.2	15.4	15.9	12.1	11.0	11.5	11.0

Table 14: Task accuracy (%).

Model	**Cond.**	**Prompt**	**DE**	**ZH**	**ES**	**AR**
GPT-5	EN→X	None	97.8	99.5	99.5	97.8
	EN→X	Explicit	98.4	98.9	99.5	98.9
	X→EN	None	94.0	95.6	94.5	96.2
	X→EN	Explicit	94.0	93.4	94.0	94.5
Claude 4.5	EN→X	None	96.7	94.0	97.3	96.7
	EN→X	Explicit	96.7	94.5	97.3	96.2
	X→EN	None	10.4	9.9	6.0	4.4
	X→EN	Explicit	9.9	9.9	5.5	3.3
Command R+	EN→X	None	91.8	89.0	95.6	80.8
	EN→X	Explicit	91.8	87.4	96.7	83.0
	X→EN	None	1.1	1.1	0.5	0.5
	X→EN	Explicit	1.1	0.5	0.5	0.5

Table 15: Language fidelity (%) by condition and system prompt.

Model	**Cond.**	**Prompt**	**DE**	**ZH**	**ES**	**AR**
GPT-5	EN→X	None	57.1	59.9	59.3	60.4
	EN→X	Explicit	58.2	59.9	58.8	56.0
	X→EN	None	55.5	50.5	49.5	54.4
	X→EN	Explicit	54.9	57.7	53.8	53.8
Claude 4.5	EN→X	None	49.5	46.7	50.5	48.9
	EN→X	Explicit	48.4	52.7	55.5	54.4
	X→EN	None	48.4	47.8	52.7	50.5
	X→EN	Explicit	48.4	51.1	51.6	48.9
Command R+	EN→X	None	15.4	13.2	15.4	15.9
	EN→X	Explicit	14.8	17.6	14.8	15.4
	X→EN	None	12.1	11.0	11.5	11.0
	X→EN	Explicit	11.5	12.6	12.6	11.0

Table 16: Task accuracy (%) by condition and system prompt.