Final Report: Human Behavioral Phenomena in LLMs: Investigating Code-Switching and Recency Bias Across Multilingual Settings

Raghav Pant
rgvpant@gmail.com

Dan Sachs
ds1731@georgetown.edu

Emilio Barkett
eab2291@columbia.edu

Supervised Program for Alignment Research — Spring 2026

Abstract

As large language models (LLMs) are deployed to a global user base, understanding how they respond to multilingual and multicultural input is a pressing alignment concern. This paper investigates two related questions: (1) how do language models exhibit language or recency bias in multilingual settings, and (2) how does the grammatical structure of code-switched input shape the language in which frontier LLMs respond – including instruction following? We begin by building a simple multilingual dataset and evlauating the biases of ten LLMs in a controlled setting, observing that while LLMs do exhibit language bias, this is linked most strongly to query language rather than order. Then, using the CodeMixQA benchmark across four languages (Hindi, Simplified Chinese, Spanish, and Urdu), we evaluate a subset of the 10 models — Claude Opus 4.6, Gemini 2.5 Flash, GPT-4.1-mini, GPT-5.4, and Qwen3.6-plus — across approximately 7,100 API calls. Our central finding is that grammatical matrix language, rather than surface token frequency alone, is the dominant driver of model response language. This grammar-forcing effect is universal across all models and language pairs tested, yet factual accuracy is unaffected, pointing to a purely metalinguistic behavior rather than a comprehension limitation. Critically, models diverge sharply in how well they override this effect under explicit instruction, with significant alignment implications for multilingual deployment.

Keywords alignment research · machine learning · AI safety

1 Introduction

The deployment of large language models is increasingly multilingual: users across the globe interact with these systems in their native tongues, and code-switching — the interweaving of two or more languages within a single utterance — is a natural feature of bilingual and multilingual communication. Despite this, the behavior of frontier LLMs in code-switched settings remains poorly understood from an alignment perspective. If a model's response language can be reliably manipulated by the syntactic structure of the input, this represents a potentially exploitable inconsistency between a developer's intent (specified via a system prompt) and actual model behavior. This paper addresses two related questions. First, does the grammatical structure of a code-switched prompt — specifically, whether English or the target language serves as the syntactic matrix — predict the language of the model's response, above and beyond the simple proportion of non-English tokens? Second, when a system prompt explicitly instructs a model to respond in English, does that instruction override the pull of the grammatical matrix? These questions are relevant to AI safety and alignment in several respects. Language choice in model outputs is a

parameter that developers and deployers alike often wish to control: a customer service deployment may require English responses regardless of how users phrase their queries. If grammatically cued code-switching can override such instructions, this represents a failure of instruction fidelity. More broadly, systematic differences in how models process structured multilingual input — differences that persist across architectures and scales — suggest that multilingual alignment is an underexplored dimension of model behavior. Our contributions are: (1) a grammar-forcing evaluation framework applied to five frontier LLMs across four languages using the CodeMixQA benchmark; (2) a parametric token-proportion control experiment disentangling grammatical structure from surface character density; (3) a system-prompt override experiment measuring instruction compliance under grammatical conflict; and (4) a cross-benchmark comparison between response-language grammar forcing and fact-selection language primacy.

2 Related Work

Diverse research exists probing whether LLMs exhibit biased behaviors in the same way that humans tend to. [1] is one such motivating work by Emilio, showing that reasoning in LLMs controls their propensity for truth-bias and sycophancy. A related question was explored in [3], where the existence of primacy and recency bias in LLMs was established via experimentation. Such questions around LLM behavior arre fundamental to understanding how to control bbiases and ensure that model alignment is responsible.

In the multilingual setting, [2] investigated the performance of chain-of-thought (CoT) reasoning across languages, showing that the benefits of reasoning and CoT are degraded significantly in the non-English setting. [5] recently conducted a large-scale study on the impact of code-switching (the practice of interweaving English and non-English languages together in sentences, common in bilingual and multilingual people). They reveal multifaceted difficulties in processing and generating code switched text, with an overall drop in benchmark performance and question answering across a varied set of topics. [4], in related work, showing that multilingual (rather than code-switched) prompting can actually improve the diversity of responses and perspectives generated by LLMs, particularly beneficial in lower-resource or non-Western cultural contexts.

These are a sample of works that motivate the two research questions we investigate here: a) does input language have a role to play in mediating recency bias in LLMs, and b) does code-switched input result in increased performance across cultural tasks, and what is the link to language in these settings?

3 Methods

As discussed we pursued a two-phase exploratory study design in order to investigate the decoupled effects of language and position on LLM biases, before extending the study to code-switched inputs (in which multiple languages are present in the input).

In Phase 1, we evaluated five frontier LLMs (Claude Opus 4.6, Gemini 2.5 Flash, GPT-4.1-mini, GPT-5.4, Qwen 3.6+) on a purpose-built factual contradiction benchmark. The dataset comprised 25 items across three difficulty levels, each containing two mutually contradictory facts and a question, translated into nine languages (English, Spanish, French, Indonesian, Hebrew, Arabic, Hindi, Urdu, Chinese). As example would be "the phone is black" and the "phone is blue", across languages, with the models then prompted to respond to the query "what color is the phone". For each item we generated all ordered ($target_lang$, $other_lang$) permutations, varying the language of Fact 1, Fact 2, and the query independently. This design decouples fact-position from language:

comparing

$$(f1 = A, f2 = B, q = X)$$

against $(f1 = B, f2 = A, q = X)$ isolates language preference from recency. An LLM-as-judge scorer attributed each response to $fact_1$, $fact_2$, or unclear using only English canonical texts, making scoring language-agnostic. While this represents a limitation of the study design, we considered it the easiest way of designing an experiment that could efficiently probe model bias.

The second phase, exploring the impact of code-switched input text on LLM behavior, selected code-switched questions in four languages (Hindi, Chinese, Spanish, Urdu) under four code-switching methods: GF-SRC (English grammatical matrix), GF-TGT (target-language grammatical matrix), random word-level switching, and selective content-word switching, based on the CodeMixQA dataset [5]. In the first stage of the experiment, we measured P(model responds in target language) across the four code-switching conditions to better understand how models might be steered towards output in the target langauge under code switching conditions. As part of this, we added random and selective conditions to test whether non-English character/token fraction, rather than grammatical structure, drives response-language choice (assessed via Spearman rank correlation). In the second stage of the experiment, we re-ran the GF-TGT portion of the experiment with an explicit English-language system prompt to determine whether grammar-forcing survives instruction override (i.e. we add a system prompt enforcing that the models should always respond in English). Statistical comparisons used two-proportion Fisher exact tests; cross-benchmark correlations used Pearson r.

Finally, we explore via data visualization whether there is any correlation between the behaviors measured around language or recency bias in the first experiment and the grammar-forcing results of the second experiment, but find no significant correlation.

4 Discussion

The grammar-forcing pattern we observe here is consistent and significant, even under such small sample size constraints. Its appearance across the diverse models and language pairs suggests that it may reflect fundamentals about how current LLMs process code-switched input. Whether this directly reflects distributional characteristics of the multilingual training data remains to be elucidated through further study. We emphasize here that larger-scale evaluation, coupled with interpretability and attribution methods, can provide further insight into the mechanisms behind this observed pattern. Further, a more rigorous experimental design, normalizing for architectural differences, tokenization variation across languages, and data distributions may offer deeper insight into the nature of this behavior across frontier LLMs.

The absence of accuracy differences across GF-SRC and GF-TGT conditions is consistent with the interpretation that grammar-forcing is a response-generation behavior rather than a sign of model comprehension. One interpretation of these results is that models may be reading code-switched questions similarly across conditions, and default to the grammatically dominant language in output. This is less interesting from an alignment standpoint; such behavior intuitively reflects a desirable and expected property of language processing and generation. That said, a deeper study into model comprehension in code-switched settings, extending the work of [5], could better quantify these properties. However, the variation in instruction compliance across models is perhaps the most practically relevant observation. We observe an apparent split between models that readily comply with an English-override system prompt and those that do not – and its correlation with pretraining data distribution is not clear. This is an area that merits deeper mechanistic study: understanding why models behave differently despite purportedly similar alignment goals can uncover fundamental patterns and circuits in how LLMs process inputs acros and in multiple languages.

Table 1: Response language by condition — selected results. Δ is the difference in target-language response rate between GF-TGT and GF-SRC conditions (percentage points). All 20 model $\times$ language cells were statistically significant (19/20 at $p < 0.001$, 1/20 at $p = 0.005$; Fisher exact test).

Model	**Language**	**GF-SRC $\rightarrow$ Target**	**GF-TGT $\rightarrow$ Target**	**Δ**
Claude Opus 4.6	Hindi	33.3%	64.0%	+30.7
Claude Opus 4.6	Simplified Chinese	8.0%	96.0%	+88.0
Claude Opus 4.6	Spanish	18.7%	93.3%	+74.7
Claude Opus 4.6	Urdu	61.3%	100.0%	+38.7
GPT-5.4	Hindi	33.3%	70.7%	+37.3
GPT-5.4	Simplified Chinese	30.7%	98.7%	+68.0
GPT-5.4	Spanish	36.0%	89.3%	+53.3
GPT-5.4	Urdu	60.0%	100.0%	+40.0
Qwen3.6-plus	Hindi	32.5%	72.3%	+39.8
Qwen3.6-plus	Spanish	26.2%	99.0%	+72.8

Future work in this direction might focus on three areas: larger and cleaner experimental designs that isolate grammatical structure from token density and script variation more precisely; systematic diversification of override prompting to better characterize the observed deviation from instruction; and extending to a broader set of languages, particularly lower-resource ones, where the effects may differ. We emphasize again the limitations here reflect the smaller scope of this introductory study; even so, having discovered some unique effects, we encourage further exploration of this work that should include more mechanistic analyses of observed phenomena under a more tightly controlled experimental setup.

5 Conclusion

This report presents preliminary evidence that the grammatical matrix language of code-switched input is a consistent predictor of frontier LLM response language independent of accuracy; further, and more relevant to alignment, we show models differ noticeably in their adherence to this behavior under an English-forcing system prompt. The last of these results has the largest implication for LLM alignment in a global context, and we emphasize that these findings should motivate further exploration of these observed patterns, both at the evaluation level and using mechanistic attribution tools. We also find that most frontier models today are well controlled with respect to language and recency biases: the majority of models show a language primacy bias linked only to the query language, with strong controls on recency bias. Limitations exist: our experiments are limited in scale, lacking mechanistic explanation, with design choices that may introduce noise (in particular, few normalizations across scripts and the use of LLM-as-a-judge evaluations). Multilingual and code-switched interactions represent an important and underexplored dimension of LLM behavior with direct relevance to alignment, and we hope this preliminary investigation helps point toward more rigorous work in this direction. Furthermore, while the evaluations here were conducted over relatively well-supported languages, these effects are already significant: further study should focus more deeply on lower-resource languages, where deeper inequalities exist, and help elucidate how knowledge transfer and alignment can help better represent the speakers of these languages in frontier LLM technology.

References

[1] Emilio Barkett, Olivia Long, and Madhavendra Thakur. Reasoning isn't enough: Examining truth-bias and sycophancy in llms, 2025.

[2] Josh Barua, Seun Eisape, Kayo Yin, and Alane Suhr. Long chain-of-thought reasoning across languages, 2026.

[3] Xiaobo Guo and Soroush Vosoughi. Serial position effects of large language models, 2024.

[4] Qihan Wang, Shidong Pan, Tal Linzen, and Emily Black. Multilingual prompting for improving llm generation diversity, 2025.

[5] Genta Indra Winata, David Anugraha, Patrick Amadeus Irawan, Anirban Das, Haneul Yoo, Paresh Dashore, Shreyas Kulkarni, Ruochen Zhang, Haruki Sakajo, Frederikus Hudi, Anaelia Ovalle, Syrielle Montariol, Felix Gaschi, Michael Anugraha, Rutuj Ravindra Puranik, Zawad Hayat Ahmed, Adril Putra Merin, and Emmanuele Chersoni. Can large language models understand, reason about, and generate code-switched text?, 2026.