German-English Code-Switching in Large Language Models

Firat Cem Aksüt[1], Stefan Hillmann[1], Pia Knoeferle[2], Sebastian Möller[1]

[1]Technische Universität Berlin, [2]Humboldt-Universität zu Berlin,

Correspondence: aksuet@campus.tu-berlin.de, stefan.hillmann@tu-berlin.de

Abstract

Code-Switching (CS) is common in multilingual communication, yet it is unclear how well current Large Language Models (LLMs) reproduce naturally occurring switching patterns. This paper studies German–English CS ("Denglisch") generated by GPT-4o-mini and LLaMA-3.3, using Reddit data from the Denglisch Corpus as a reference. Model outputs are compared to authentic posts using established CS metrics (M-Index, I-Index, CE-SAR), an analysis of Shared Lexical Items (SLIs) as switch triggers, and a human evaluation of Perceived Naturalness and Perceived Fluency. Both models approximate global CS characteristics but differ in the diversity and complexity in comparison to real data. LLaMA-3.3 more closely matches corpus-level metrics, whereas GPT-4o-mini produces more conservative switching that is rated as significantly more natural and fluent. In addition, GPT-4o-mini reproduces SLI-triggered switching patterns similar to those found in authentic data, while this effect is weaker for LLaMA-3.3.

1 Introduction

Large Language Models (LLMs) such as GPT-4 are increasingly being used in everyday applications, such as chatbots, translation systems and writing assistants. Since the release of ChatGPT in 2022, they have quickly become widespread (Hu, 2023). At the same time, multilingualism is commonplace in digital communication spaces. According to estimates, around 3.5 billion people speak at least two languages (Grosjean, 2021). A key phenomenon here is Code-Switching (CS), i.e. switching between two or more languages within a conversation, sentence or even a word (Riehl, 2019a). CS fulfils both linguistic economy and identity-related functions, especially in informal contexts and social media (Riehl, 2019b).

Despite their strong multilingual capabilities, there has been limited research into how well modern LLMs actually replicate such naturally occurring CS patterns. Studies show that even large multilingual LLMs often lag behind specialised systems in CS-related tasks and are 'not (yet) Code-Switchers' (Zhang et al., 2023). At the same time, linguistic studies suggest that so-called *Shared Lexical Items (SLIs)*, i.e. words that occur in both languages, can trigger language switching in a targeted manner by increasing the probability of a switch in their environment (triggering hypothesis) (Broersma and de BOT, 2006).

Within this context, this paper examines the extent to which current LLMs can realistically imitate German-English CS from Reddit posts. Based on the Denglisch Corpus (Osmelak and Wintner, 2023), we have two models (GPT-4o-mini and LLaMA-3.3) generate new Denglisch posts and compare them with authentic posts using established CS metrics, an analysis of SLIs and a human evaluation.

Specifically, this paper addresses the following research questions:

- **Research Question (RQ) 1:** To what extent can current LLMs reproduce the global CS patterns of Denglisch Reddit posts?
- **Research Question (RQ) 2:** Do LLMs exhibit SLI-triggered CS patterns similar observed to those in real Denglisch data?

2 Background

2.1 Code-Switching

Code-switching (CS) generally refers to the switch between two (or more) languages within a single utterance, sentence, or conversation (Riehl, 2019a; Muysken, 2011). The matrix language, which determines the grammatical structure and the main part of the expression, is supplemented by the embedded language, which inserts individual words, phrases, or expressions from another language

Proceedings of the Thirteenth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 87–100

(Myers-Scotton, 2017). The phenomenon typically occurs among bilingual speakers and multilingual communities and requires a corresponding level of linguistic competence. Typically, the grammatical structures of the languages are not violated. CS is influenced by sociocultural, situational, or contextual factors. Poplack (1980) and Kootstra (2015) provide an psycholinguistic view on it.

The literature distinguishes between different types of CS. Traditionally, a distinction is made between **intra-sentential CS**, **inter-sentential CS** and **tag switching** (Poplack, 1980). In intra-sentential CS, the language switch occurs within a sentence, typically at syntactically compatible boundaries. In inter-sentential CS, the switch occurs between sentences or larger units of speech. Tag switching, on the other hand, refers to the insertion of short fixed expressions (e.g. discourse markers) from another language. In addition, **intra-word CS** (Myers-Scotton, 1989) is often described, in which the switch occurs within a word (e.g. through the combination of morphemes).

2.2 Code-Switching Metrics

In order to quantify CS, there are a number of established measures that capture different aspects of language switching. The metrics used in this study are primarily the M-Index, I-Index and CSR.

The normalized (Chi et al., 2024) **M-Index** (Mixing Index) (Barnett et al., 1999, p. 202)measures the balance of the languages involved based on relative token frequency. Values close to 0 indicate a strong dominance of one language, while values close to 1 indicate an even distribution.

The **I-Index** (Integration-Index) quantifies the density of switch points in the text (Guzmán et al., 2017). It measures how often two neighbouring tokens are assigned to different languages. A value of 0 means no language switch, while a value of 1 means a switch occurs at every token.

The **CSR** (Code-Switching Rate) corresponds to a simplified form of the I-Index. It describes the proportion of token boundaries at which a language change occurs and can be interpreted as the probability of a change at a random position.

Finally, **CESAR** measures the degree of CS relative to a reference language. It combines the occurrence, i.e. the proportion of tokens that contain tokens from another language, with the burstiness, i.e. how strongly these foreign-language tokens occur in clusters within the units (Abidi and Smaïli, 2021). A value of 0 means that the text is exclusively in the reference language, while a value of 1 means that it does not occur at all.

2.3 Shared Lexical Items

The *triggering hypothesis* (Clyne, 2003) assumes that certain lexical elements that occur in both languages of a speaker's repertoire increase the probability of a language switch. These 'shared lexical items' (SLIs) include proper names, culturally specific terms, and terms for which there is no direct translation. The presence of such a trigger in an utterance appears to cause the brain to switch from one language to another (BROERSMA and DE BOT, 2006).

Wintner et al. (2023) confirm this assumption in a large-scale analysis of several language pairs. They have demonstrated that the probability of a switch increases significantly when an SLI appears shortly before the switch and is slightly lower when it occurs afterwards. They distinguish SLIs according to origin, such as shared English, shared German or shared other. These findings form the basis for the present study, in which we examine whether modern LLMs show similar SLI-triggered patterns in German-English CS.

2.4 LLMs and Code-Switching

Although modern LLMs have strong multilingual capabilities, they still show limitations in CS. Studies show that even large models often lag behind specialised systems in CS-related tasks (Zhang et al., 2023). Challenges arise in particular from correct token-level language identification and other token-level processing, dealing with ambiguous word forms, and adhering to grammatical CS patterns (Çetinoğlu et al., 2016). For the German-English language pair in particular, both CS research in general and work on LLM behaviour under CS are still relatively underrepresented and only few annotated resources are available. Osmelak and Wintner (2023) address this gap by introducing the Denglisch Corpus, a German-English CS corpus that serves as an empirical basis for our study.

3 Data and Methodology

3.1 Denglisch Corpus

For our experiments, we use the Denglisch Corpus of Osmelak and Wintner (2023). This is an extensive corpus of German-English CS posts from Reddit. The corpus was created specifically to reflect

naturally occurring Denglisch in informal online texts and contains around 31,500 posts. It contains various forms of CS, such as intra/inter-sentential, tag-based and hybrid forms of CS. The posts come from different subreddits and cover a wide range of topics, making the corpus a representative basis for the analysis of mixed-language online communication.

To generate and compare the LLM responses, we take a random sample of 50 posts (Input-Data) that represent the entire corpus as well as possible in terms of length, CS characteristics and stylistic variation. There are several reasons for limiting the sample to 50 examples. On the one hand, processing very large amounts of data can lead to instability (Bender et al., 2021). At the same time, a small, carefully selected input allows for a more precise analysis of model behaviour, especially with regard to few-shot scenarios.

3.2 Annotation

Osmelak and Wintner (2023) also provide a fine-grained annotation scheme and word-level classifier as part of their work. In addition to clear language labels (German/English), the scheme also includes categories for loanwords, ambiguous forms and mixed constructions, which enables precise identification of CS points. Part of the corpus was annotated manually and used to develop a classifier, which was then applied to the complete data set. They report an average tagging accuracy of 0.764 at the sentence level (i.e., across all tags within a sentence), and, for instance, F_1 scores of 0.96 and 0.97 for German and English tags, respectively (Osmelak and Wintner, 2023, Table 5).

For our experiments, we reuse this classifier to consistently annotate all LLM-generated texts. To make the labels comparable across corpus and model outputs, we map the original fine-grained tags onto a compact tagset with four language categories. All German-specific labels are collapsed into a single GERMAN class, all English-specific labels into ENGLISH, mixed labels (e.g. hybrid forms) into MIX, and discourse marker, interjections and similar neutral items into a TAG/OTHER class. In addition, we exploit the orignal shared-item labels of the classifier. All tokens tagged as shared (e.g. shared English, shared German, shared other) are marked with an SLI flag and their origin. This annotation allows us to automatically compute all CS metrics and to identify both different CS types and SLI-triggered switches across the three sources.

3.3 Models

We compare the GPT-4o-mini and LLaMA-3.3-70B-Instruct (LLaMA-3.3) models. The selection is based on criteria for CS evaluation scenarios. These include multilingualism, model size, accessibility, and usage costs. According to Zhang et al. (2023), also the MMLU[1] benchmark is suitable for this purpose although it is not explicity designed as a CS benchmark.

While GPT-4o-mini is a proprietary API model with paid token rates and an MMLU score of 82 % (OpenAI, 2024), LLaMA-3.3 is a freely usable open-source model that can be run locally or in cloud environments and an MMLU score of 86 % (Hugging Face, 2024). Furthermore, the models differ in parameter count (GPT-4o-mini: 8B[2]; LLaMA-3.3: 70B) and training methods. Since model size and training regime have been shown to influence the quality of generative tasks (Brown et al., 2020), this model comparison provides cross-validation of the CS patterns across parameter and training differences.

3.4 Generation Setup

To generate the LLM outputs, we use a standardised prompt design that is identical for both models. The prompts given to the LLMs and discussed here are provided in Appendix A. Each of the 50 input posts is given to the respective models together with a brief introduction to CS. The respective CSR is also provided with each of the sample posts so that the models can recognise the typical CS intensity of the data. Our pilot tests with and without this additional information showed a positive affect on the models' performance regarding CS generation. The structure of the prompt follows the few-shot principle, which has proven effective for controlling linguistic generation tasks (Brown et al., 2020).

In the instructions, the models are asked to generate a 3-sentence Reddit post that contains natural German-English CS and varies in content and style. In addition, the generated post should have a CSR close to the median value of the examples provided.

[1] MMLU (Massive Multitask Language Understanding) is a multiple-choice test that evaluates LLMs in 57 different subject areas to measure their cross-domain knowledge coverage and problem-solving skills.

[2] A pre-print article (https://arxiv.org/abs/2412.19260) discloses the parameter size (8B), the final article does not include absolute numbers (Ben Abacha et al., 2025).

To promote stylistic diversity, we use a temperature of 0.9. Lower values led to predominantly monolingual or repetitive responses in pilot runs. All generations are performed under identical conditions for both models. The outputs are then saved and further processed for annotation and metric calculation. This controlled setup mitigates prompt bias and found differences in the CS patterns are can be attributed to the models themselves.

3.5 Human Evaluation Setup

To supplement the automatic metrics, we conducted a human evaluation to assess the Perceived Naturalness and Perceived Fluency of the posts. The survey was implemented via a crowdsourcing platform and targeted participants located in the German-speaking DACH region (Germany, Austria, Switzerland). Eligibility criteria required participants to have German as a (co-)native language and to report at least intermediate proficiency in English, ensuring that they were able to understand and evaluate German–English CS posts.

After excluding two participants based on a control question, we could include responses of 133 individuals in our analysis. The age distribution is predominantly in the 25–44 age range. Younger (under 18) and older participants (over 65) are clearly under-represented. The vast majority of participants stated German as their native language ($\approx 91\%$), with only a few naming English or other languages as their native language. Among the German-speaking participants, more than three-quarters rated themselves as 'advanced' or 'fluent' in English, suggesting a high level of bilingual competence and good conditions for assessing German-English CS patterns.

Each participant completed one questionnaire containing five Reddit posts with German–English CS, randomly sampled from the three sources GPT-4o-mini, LLaMA-3.3, and the Input-Data. In total, 14 posts per source (42 posts overall) were included in the evaluation, each of which was rated 15 times, resulting in 630 individual ratings. The posts were randomly assigned to questionnaire versions and randomised in order to reduce position and anchoring effects. The exact wording of all questions and items discussed here and in the following are provided in Appendix B.

The central scales for the assessment were Perceived Naturalness and Fluency. Each of them were measured on a five-point Likert scale ranging from 'very unnatural/unsmooth' (1) to 'very natural/smooth' (5). After the naturalness rating, participants who selected 'very unnatural' or 'rather unnatural' were presented with a follow-up question asking to select from pre-defined reasons why the post seemed unnatural (e.g. unnatural mixture of German and English, unusual word choice, grammar errors, lack of natural flow). Conversely, participants who selected 'very natural' or 'rather natural' received an analogous question asking for reasons why the post seemed natural (e.g. authentic language switching, natural mixture of German and English, typical Reddit phrases, natural word choice). If 'neutral' was selected, no follow-up question on reasons was shown.

In addition, the questionnaire collected demographic information (age group, native language, self-assessed proficiency in the other language) as well as self-reported usage patterns. This means, how often participants read and post in online forums such as Reddit, and how often they themselves use CS in online texts. These variables serve to contextualise the ratings and to identify potential factors that might influence the perception of CS.

The collected data was then evaluated descriptively. We compared the mean values of the evaluations across models and corpora and examined whether the Perceived Naturalness or Fluency correlated with objective CS metrics (CSR, M-Index, I-Index, CESAR).

4 Quantitative Results

4.1 Global CS Patterns

In a first step, we examine the global properties of the CS patterns generated by the models using the three metrics presented (see subsection 2.2). The M-Index describes the linguistic diversity within a post, the I-Index describes the frequency of language changes, and CESAR describes the complexity and burstiness of the CS patterns.

As shown in Figure 1, LLaMA-3.3 achieves higher mean values than GPT-4o-mini (0.247 vs. 0.196) in the M-Index, which indicates greater linguistic diversity in the LLaMA-3.3 responses. However, the Input-Data show higher values (0.369), meaning that none of the models reach the level of real Denglish posts. A similar picture emerges for the I-Index. With a mean of 0.165, LLaMA-3.3 is clearly above GPT-4o-mini (0.099) and the difference to the Input-Data (0.174) is remarkable smaller. LLaMA-3.3 thus tends towards more frequent language changes and a more inte-

grated language mix overall, while GPT-4o-mini remains more conservative.

The differences between the two models are less pronounced in the CESAR score. GPT-4o-mini achieves a mean of 0.273, while LLaMA-3.3 achieves slightly higher values of 0.283. However, none of both match the CESAR score of the Input-Data (0.336), indicating that neither the complexity nor the burstiness of the CS patterns in the corpus are fully replicated. Figure 1 shows that LLaMA-3.3 is closer to the corpus mean values across all metrics than GPT-4o-mini. Overall, these results only partially support Hypothesis 1. Both models imitate global properties of the CS structure, but underestimate the degree of linguistic diversity, alternation frequency and complexity, with LLaMA-3.3 approximating the Input-Data more consistently than GPT-4o-mini.

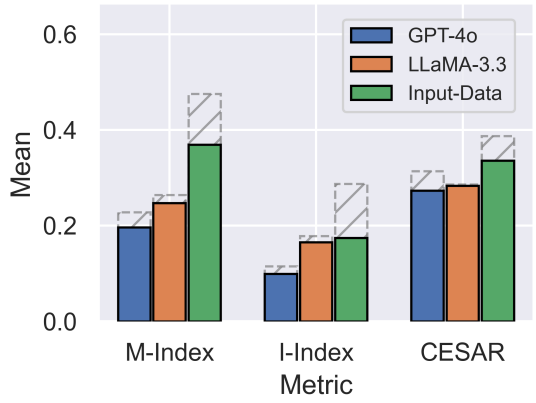

Figure 1: Mean values and standard deviation of M-Index, I-Index, and CESAR for GPT-4o-mini, LLaMA-3.3, and the Input-Data.

4.2 SLI-triggered CS

Building on the concept of Shared Lexical Items (SLIs) and the triggering hypothesis from subsection 2.3, we now examine whether SLIs occur preferentially near CS points in our data (Hypothesis 2). Specifically, we examine the extent to which LLMs reproduce the SLI trigger effects observed in the Input-Data.

For each of the three sources (GPT-4o-mini, LLaMA-3.3, Input-Data), we construct a 2×2 contingency table that compares SLIs vs. non-SLIs and 'near to CS' vs. 'not near to CS'. A token is considered to be near to CS if a switch point occurs within a range of six words. This range captures both upstream and downstream trigger effects. We use the corresponding frequencies to calculate the Relative Switching Propensity (RSP), i.e. the ratio of the CS probability near SLIs to the CS probability near non-SLIs. Values greater than 1 indicate a positive correlation between SLIs and CS. The model values are averaged over 50 generation runs (2,500 posts per model), with the Input-Data corresponding to the 50 original posts. The central results are summarised in Table 1.

	GPT-4o-mini	**LLaMA-3.3**	**Input-Data**
% SLIs near CS	63.00	87.50	67.06
% non-SLIs near CS	32.81	66.17	37.84
RSP	1.92	1.32	1.77

Table 1: Shared lexical items (SLIs) and their association with CS: percentage of SLIs and non-SLIs near CS and relative switching propensity (RSP) for GPT-4o-mini, LLaMA-3.3, and the Input-Data.

The Input-Data shows a clear SLI trigger effect. Of the 85 SLIs in the sample, 57 occur near CS (67.06 %), while this is only true for 1,077 of 2,845 non-SLIs (37.84 %). This results in an RSP of 1.77. This means that CS is 1.77 times more likely to be near an SLI than near a non-SLI. For the Input-Data, Fisher's exact test shows that SLIs occur near CS significantly more often than non-SLIs ($p < 0.05$).

GPT-4o-mini clearly reflects this pattern. On average, 63.00 % of SLIs are close to CS, while this applies to only 32.81% of non-SLIs. The resulting RSP is 1.92, which is slightly above the value of the Input-Data (1.77). For GPT-4o-mini, SLIs occur near CS significantly more often than non-SLIs according to Fisher's exact test ($p < 0.05$).

LLaMA-3.3 shows a slightly different profile. Although 87.50 % of SLIs are close to CS, the proportion of non-SLIs close to CS is also significantly higher at 66.17 %. Accordingly, the RSP is lower at 1.32, as the baseline CS probability is also high for non-SLIs. For LLaMA-3.3, Fisher's exact test does not indicate a significant association ($p \geq 0.05$).

Overall, Input-Data and GPT-4o-mini exhibit a clear SLI trigger effect with comparable RSP values, whereas LLaMA-3.3 shows a weaker association, reflecting a generally higher CS rate across tokens. From a quantitative perspective, the results thus provide partial confirmation of Hypothesis 2. SLIs are associated with an increased CS probability, and GPT-4o-mini reproduces this pattern closer to the Input-Data than LLaMA-3.3.

4.3 Human Evaluation

The results of the automated analysis were supplemented by a human evaluation, in which the Perceived Naturalness and Fluency of the posts were assessed.

To assess **Perceived Naturalness**, participants rated all posts on a five-point Likert scale from 'very unnatural' (1) to 'very natural' (5). Overall, GPT-4o-mini was rated significantly more positively than LLaMA-3.3 and the Input-Data. The categories 'rather natural' or 'very natural', were selected in around 47 % of the ratings for GPT-4o-mini fall into the categories 'rather natural' or 'very natural', compared to 33 % for LLaMA-3.3 and 30 % for the Input-Data. Conversely, the Input-Data was most frequently rated as (rather) unnatural (56 %), followed by LLaMA-3.3 (50%) and GPT-4o-mini (36 %). Interestingly, in terms of mean scores, GPT-4o-mini achives the highest Perceived Naturalness (3.19), followed by LLaMA-3.3 (2.77) and the Input-Data (2.56). The standard deviations are very similar across all three sourced at around 1.2 points on the five-point scale. Mann-Whitney U tests with Bonferroni correction indicate that the distributions of naturalness ratings differ significantly between all three sources (corrected $p < 0.01$).

The **Perceived Fluency** of the language switches was also measured on a five-point scale from 'very unsmooth' (1) to 'very smooth' (5). Here, too, the LLMs performed better overall than the Input-Data. For GPT-4o-mini, about half of the ratings fall into the top two categories ('rather fluent' and 'very fluent'), while for LLaMA-3.3 the proportion is slightly lower, and the Input-Data is perceived as (rather) non-fluent significantly more often. The mean fluency scores again show a clear ordering. GPT-4o-mini obtains the highest mean rating (3.29), LLaMA-3.3 scores 3.02 and the Input-Data 2.60, with standard deviations close to 1.2 for all three sources. Mann-Whitney U tests indicate significant differences between both LLMs and the Input-Data (corrected $p < 0.001$), while the difference between GPT-4o-mini and LLaMA-3.3 in fluency ratings is not statistically significant. Figure 2 summarises the mean ratings for both Perceived Naturalness and Fluency and illustrates the consistent advantage of GPT-4o-mini over LLaMA-3.3 and the Input-Data.

The open-ended responses regarding the reasons for (un)naturalness provide additional insights into the perception of the models. The most frequently cited reason for unnaturalness is an 'Unnatural mixture of German and English'. This reason occurs most frequently with GPT-4o-mini, followed by LLaMA-3.3 and the Input-Data. 'Unusual choice of words' is mentioned with similar frequency in all three groups. It is striking that a 'Lack of natural speech flow' is mentioned particularly frequently in LLaMA-3.3 and the Input-Data, while this reason appears less frequently in GPT-4o-mini. Grammatical errors, on the other hand, are mainly attributed to the Input-Data and significantly less frequently to the LLM outputs, which suggests that the models usually produce relatively clean grammar.

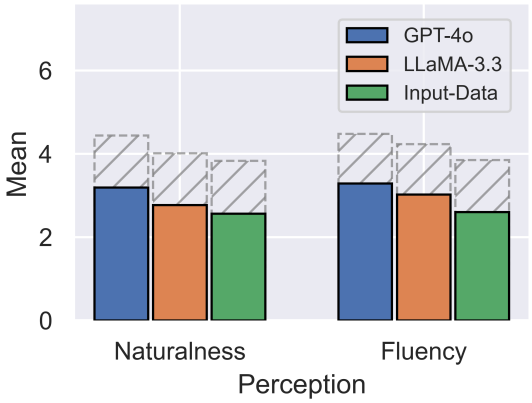

Figure 2: Mean values and standard deviation of Perceived Naturalness and Perceived Fluency of GPT-4o-mini, LLaMA-3.3, and the Input-Data (5-point Likert scale).

To qualitatively illustrate the differences between GPT-4o-mini, LLaMA-3.3, and the Input-Data, Table 2 shows representative three-sentence posts from each source used in our experiments. For each source, we show one post with low, medium and high CS rates to illustrate variation across CS intensity levels.

Overall, and considering only the numerical results, the human evaluation indicates that the two LLMs employed, with —GPT-4o-mini in particular, outperform the authentic Reddit posts with respect to perceived naturalness and fluency of language. We attribute this primarily to the models' capacity to produce code-switching, combined with their tendency to generate text that adheres closely to standard grammatical and orthographic conventions, in contrast to typical Reddit contributions. Had the LLMs been able to systematically imitate the non-standard, error-prone style characteristic

GPT-4o-mini	LLaMA-3.3	Input-Data
Post 1 (low CS)		
Die letzten Tage waren einfach *crazy*. Ich habe so viel Zeit damit verbracht, mich durch die neuen Marvel-Serien zu *binge-watchen* und es ist als ob ich in einer anderen Welt bin. Die Storylines sind so gut, dass ich ständig über die Charakterentwicklung nachdenke und dabei dachte ich: *Are they really going to pull this off?* Hoffentlich geht das nicht in die Hose wie bei manchen Filmen.	Das Problem mit den neuen Datenschutzbestimmungen ist, dass sie oft sehr komplex sind und man sich durch die kleinen Druckerzeichen und den *fine print* kämpfen muss, um zu verstehen, was man da eigentlich akzeptiert, und ich denke, das ist ein bisschen wie ein *Catch-22*, weil man entweder seine Privatsphäre opfern oder sich durch einen *terms of service* kämpfen muss, der länger ist als ein Harry-Potter-Buch.	*There is no glory in prevention* ja, aber ob die absolute Anzahl der Kontakte durch eine Ausgangssperre wirklich reduziert wurde, ist schon zu hinterfragen.
Post 2 (medium CS)		
Das Wetter hier ist wirklich verrückt. Letzte Woche waren wir bei 30 Grad und jetzt liegt Schnee. *Just when I thought*, wir könnten endlich den Frühling genießen, kommt der Winter zurück. Ich meine, *seriously*, was soll das? Es ist, als würde *Mother Nature* mit uns spielen.	Ich bin gerade total *frustrated* mit dem aktuellen Stand der Dinge, es ist einfach *too much*, wenn man sieht, wie die Politiker mit dem Thema Umsatzsteuer umgehen, *total unfair*, und ich denke, es ist Zeit für einen *change*, vielleicht sollte man einfach das System hinterfragen und nach neuen Lösungen suchen, wie zum Beispiel eine *flat tax*, aber das ist natürlich ein komplexes Thema, das mehr diskussion und expertenwissen braucht, um es richtig zu verstehen und umzusetzen.	Amerikanische *mandatory sentences* und auch *three strike laws*, grade in kombination mit der todesstrafe, sind extremst menschenunwürdig und von der Art und Weise wie Menschen zu unfairen *plea-deals* genötigt werden weil das justizsystem einfach komplett zusammen brechen würde wenn auch nur 40% aller Menschen auf ne *trial* bestehen würden will ich gar nicht erst anfangen.
Post 3 (high CS)		
Ich habe neulich eine Doku über die amerikanische Kultur gesehen und *wow*, die Unterschiede sind echt krass. In den *USA* feiern sie *Thanksgiving* mit *Turkey* und *Family*, während wir hier in Deutschland einfach unser Weihnachtsessen haben. Das klingt fast wie ein Wettbewerb, welcher Feiertag die meisten Kalorien bringt. Und dann die *Black Friday Sales – what's up with that*? Ich kann mir nicht vorstellen, dass wir hier so etwas machen würden.	Ich habe gerade ein *super cooles* Video auf *YouTube* gesehen, das war ein totaler *game changer*, weil es zeigte, wie man mit einem einfachen hack das Leben leichter machen kann, zum Beispiel mit einem *life hack*, um den Alltag zu vereinfachen, und ich dachte, das wäre ein interessanter *topic* für eine Diskussion, weil es zeigt, dass man mit einem *little trick* und einem bisschen *creativity* ein *lot* erreichen kann.	das problem dabei ist, wenn man *browser* erweiterungen wie *Reddit Enhancement Suite* benutzt steht hinter jedem namen den man jemals *ge-up* oder *downvoted* hat die summe der *votes*. Deine art zu *voten* ist dann wie wenn du auf *youtube* einmal nen *rap* video *upvotest* weil der text besonders kreativ ist oder aus welchem grund auch immer obwohl du *rap* musik hasst und dann hast du die nächsten 2 monate auf deiner *youtube* startseite nur noch *rap* videos / kanäle.

Table 2: Representative three-sentence posts from GPT-4o-mini, LLaMA-3.3, and the Input-Data illustrating typical German–English CS patterns observed in our experiments. Posts are ordered from low to high CS rate.

of many Reddit posts, it is likely that the evaluators would have judged the LLM-generated content less favorably than the original dataset messages.

A further perspective on these findings arises from the composition of our evaluator sample. The participants are not representative of the general population of German Reddit users who frequently engage with the platform. This limitation and its implications are discussed in more detail in the following section.

Crucially, the primary aim of the human evaluation was not to assess the LLMs' ability to perfectly replicate prototypical Reddit texts—or, more broadly, highly irregular user-generated content—but rather to determine whether the models can produce code-switching in a manner that human readers find convincing.

5 Discussion

This study examined the extent to which current LLMs can realistically imitate German-English CS in Reddit-like texts. The focus was on global CS patterns in comparison to the Denglisch Corpus (RQ 1) and SLI-triggered language switching (RQ 2). This was supplemented by a human evaluation of Perceived Naturalness and Fluency. In the following, we discuss the findings in terms of these questions and possible technical causes.

With regard to RQ 1, the models only partially

reproduce the global CS characteristics. In terms of the M-Index and I-Index, LLaMA-3.3 shows smaller deviations from the Denglisch corpus than GPT-4o-mini, indicating a higher degree of linguistic mixing and more frequent language switching. For CESAR, both models exhibit similarly small differences relative to the Input-Data. From a purely metric point of view, LLaMA-3.3 initially appears to be more 'English-like' than GPT-4o-mini. However, the human evaluation shows that greater mixing and higher alternation frequency are not automatically perceived as more natural or fluent. GPT-4o-mini achieves significantly higher ratings for Perceived Naturalness and Fluency than LLaMA-3.3 and the Input-Data, even though it switches more conservatively overall. This supports the assumption that natural Denglisch is characterised more by selective, well-embedded language switching than by 'as much CS as possible'. This is a pattern that GPT-4o-mini realises more strongly than LLaMA-3.3.

RQ 2 concerned SLI-triggered language switching. Here, the Input-Data and GPT-4o-mini show very similar patterns. In both cases, the probability of a switch in the vicinity of SLIs is significantly increased. The RSP in these cases is in a comparable range. Although LLaMA-3.3 also has an RSP value above 1, the effect is weaker and statistically not significant. Considering the triggering hypothesis, this can be interpreted as GPT-4o-mini apparently having learned to use certain shared lexemes as natural 'anchor points' for language switching, while LLaMA-3.3 uses CS in a much more widespread manner. Due to the overall high CS rate, the switch probability also increases in the vicinity of non-SLIs, which weakens the relative trigger effect. RQ 2 is thus also partially confirmed. SLIs clearly act as triggers in both the Input-Data and the GPT-4o-mini outputs, but this mechanism is significantly weaker in LLaMA-3.3.

The observed differences can be plausibly explained by technical and linguistic factors. GPT-4o-mini belongs to the GPT-4 family and was trained on a broad, multilingual database, followed by an *Reinforcement Learning from Human (RLHF)* phase with a strong focus on dialogical and user-oriented scenarios (OpenAI, 2024). OpenAI states that GPT-4 shows significantly improved performance in non-English languages compared to previous versions (OpenAI, 2024). Given this background, the strong performance of GPT-4o-mini in a German-English context can be understood as a result of diversified training. In addition, RLHF tuning may have contributed to GPT-4o-mini tending to respond in the language of the input and only changing languages when stylistically or contextually motivated, as indirectly evidenced by the studies of (Zhou et al., 2024). This assumption is supported by the results of this work regarding the selective use of CS in GPT-4o-mini outputs.

LLaMA-3.3-70B-Instruct is based on Meta's LLaMA-3 architecture, which was pre-trained on approximately 15 trillion publicly available tokens (Grattafiori et al., 2024). According to the model card, the corpus contains texts in over 30 languages but is heavily dominated by English. Only about 5 % of the training data is non-English content (Hugging Face, 2024), so the proportion of explicitly German texts is probably well below 1 %. Instruction fine-tuning was also predominantly carried out on English-language prompts (Hugging Face, 2024). This configuration is consistent with the findings of Meeus et al. (2024) and Chang et al. (2023), which show that models trained primarily in English and highly multilingual systematically lose quality in underrepresented languages ('curse of multilinguality'). This offers a plausible explanation for the weaker performance of LLaMA-3.3 in the Denglisch context.

In substance, this has several consequences for the evaluation of CS in LLMs. First, it becomes clear that surface metrics such as M-Index, I-Index and CESAR provide important information, but without human evaluation and psycholinguistically motivated measures such as RSP, they can easily be misleading. One model may be closer to the Input-Data in terms of metrics, yet appear less natural than another. Second, the SLI findings show that LLMs can internalise subtle, cognitively motivated patterns of bilingual language production under suitable training conditions. GPT-4o-mini closely replicates the SLI trigger effect of the Input-Data, while LLaMA-3.3 only replicates it to a lesser extent. Thirdly, the diverging profiles of GPT-4o-mini and LLaMA-3.3 underscore that CS competence is highly model-dependent and that statements about the CS ability of LLMs must always be interpreted against the background of specific training data and fine-tuning regimes.

6 Conclusion

This study examined the extent to which current LLMs can replicate German-English CS in Red-

dit posts. Compared to the Denglisch Corpus, both models show that they only approximate global CS patterns. LLaMA-3.3 is closer to the Input-Data values from the corpus in terms of M-Index and I-Index, while GPT-4o-mini is perceived as significantly more natural and fluent in the human evaluation.

With regard to SLI-triggered CS, GPT-4o-mini shows a pattern that is very similar to the Input-Data, while the relative trigger effect is significantly weaker in LLaMA-3.3. Overall, this suggests that LLMs are indeed capable of internalising more refined, psycholinguistically motivated CS structures, but that their expression is highly model-dependent.

7 Future Work

This study provides relevant insights into the ways in which LLMs emulate German–English code-switching. At the same time, the identified limitations and the obtained results suggest several avenues for further research that can be pursued in subsequent studies. In addition to the question of the extent to which LLMs are capable of generating irregular user-generated content, the following code-switching-related topics may be addressed:

- By fine-tuning selected LLMs on annotated CS corpora, it is possible to investigate whether the models generate more realistic language switches as a result. This would demonstrate the extent to which CS can be learned and controlled by LLMs.
- A comparison of different language combinations, such as Turkish-German, can provide information about whether the observed model behaviour is stable across languages or specific to the German-English language pair. This would allow for a better differentiation between language pair- and model-specific influences.
- The use of alternative corpora, such as spoken dialogue data or other platforms beyond Reddit, could provide deeper insights into how the characteristics of the Input-Data influence CS behaviour in LLMs.
- A focus on prompting could help to understand the extent to which this affects CS behaviour or what remains shaped by internal training patterns.
- The development of a standardised benchmark for evaluating CS in LLMs would create a basis for reproducible research and enable systematic model comparisons.
- Finally, it could also be investigated to what extent the models use CS in a discourse-specific manner. This means, for example, whether they use CS for emphasis, group identification, or to mark contrast.

Limitations

Like any study, this work has several limitations that should be taken into account when interpreting the results. These mainly concern the data basis, the selection of models, the analysis tools used, and the design of the human evaluation. We consider the following the most important.

One key limitation concerns the data used. The Denglisch Corpus is based on Reddit posts, which reflect authentic and informal language use but represent a very specific type of Code-Switching. This is characterised by written online communication, platform culture and specific socio-demographic user groups. Verbal Code-Switching in everyday life, for example in spontaneous conversations, can be significantly different and was not taken into account in this work.

In addition, the data set could partially be considered outdated. Some of the subreddits analysed originate from earlier years (e.g. de_2013), which may have a negative impact on the representativeness of the data. The limited scope of the Input-Data used also constitutes a limitation since only a small, randomly selected portion (50 posts) of the total Denglisch Corpus was used. As a result, extreme cases and marginal phenomena of Code-Switching may have been under-represented.

Another limitation concerns the annotation of the data. Automatic categorisation using the CRF-based classifier carries the risk of misclassification, especially in the case of complex or ambiguous words. Since many evaluations in this work are based on these annotations, even minor errors in labeling can have an impact on the calculated metrics. For practical reasons, it was not possible to perform a complete manual check in the context of this work.

The transferability of the results is also limited in terms of content. The study only examined the German–English language pair. However, Code-Switching is a language-specific and culturally in-

fluenced phenomenon that can vary greatly in structure, frequency, and motivation in other language constellations. The findings obtained here cannot, therefore, be easily transferred to other language pairs or communication contexts.

Finally, the selection of LLMs is also limited. Only two models were used in the evaluation. They differ in terms of architecture, training data, and licensing, but do not represent the full range of the existing LLM spectrum. The results, therefore, cannot be generalized to other model classes.

Ethical Considerations

The study raises no ethical concerns regarding data collection or analysis, as it relies exclusively on publicly available Reddit data and anonymized, automatically generated text, without involving personal or sensitive information. A potential ethical risk lies in the possible misuse of the findings to generate synthetic posts for online forums such as Reddit, which are increasingly difficult to distinguish from human-authored content. Such applications could facilitate the spread of misinformation or manipulative content. At the same time, more realistic and context-sensitive language generation can also have positive societal effects, for example, by enabling more inclusive and natural interactions in advisory, support, or recommendation scenarios. We therefore emphasize that the results should be interpreted and applied with care, and that responsible use of code-switching–capable language models is essential.

All participants in the human evaluation of the generated and Reddit data used a crowdsourcing platform, participated voluntarily, and were reimbursed at least the German minimum wage.

References

Karima Abidi and Kamel Smaïli. 2021. CESAR: A new metric to measure the level of code-switching in corpora -Application to Maghrebian dialects. In *Springer series "Advances in Intelligent Systems and Computing"*, Springer series "Advances in Intelligent Systems and Computing", Amsterdam, Netherlands.

Ruthanna Barnett, Eva Codó, Eva Eppler, Montse Forcadell, Penelope Gardner-Chloros, Roeland van Hout, Melissa Moyer, Maria Carme Torras, Maria Teresa Turell, Mark Sebba, Marianne Starren, and Sietse Wensing. 1999. The lides coding manual: A document for preparing and analyzing language interaction data. Technical Report Version 1.1, Tilburg University, Tilburg, The Netherlands.

Asma Ben Abacha, Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, and Thomas Lin. 2025. MEDEC: A benchmark for medical error detection and correction in clinical notes. In *Proc. ACL 2025*, pages 22539–22550, Vienna, Austria. ACL.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Mirjam Broersma and Kees de BOT. 2006. Triggered codeswitching: A corpus-based evaluation of the original triggering hypothesis and a new alternative. *Bilingualism: Language and Cognition*, 9(1):1–13.

MIRJAM BROERSMA and KEES DE BOT. 2006. Triggered codeswitching: A corpus-based evaluation of the original triggering hypothesis and a new alternative. *Bilingualism: Language and Cognition*, 9(1):1–13.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Özlem Çetinoğlu, Sarah Schulz, and Ngoc Thang Vu. 2016. Challenges of computational processing of code-switching. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 1–11, Austin, Texas. Association for Computational Linguistics.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2023. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. *Preprint*, arXiv:2311.09205.

Jie Chi, Electra Wallington, and Peter Bell. 2024. Characterizing code-switching: Applying linguistic principles for metric assessment and development. In *Proceedings of Interspeech 2024*, Proceedings of Interspeech, pages 7–11. ISCA. The 25th Interspeech Conference, Interspeech 2024 ; Conference date: 01-09-2024 Through 05-09-2024.

Michael Clyne. 2003. Dynamics of language contact: English and immigrant languages. *Dynamics of Language Contact: English and Immigrant Languages*, pages 1–282.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

François Grosjean. 2021. *Life as a Bilingual: Knowing and Using Two or More Languages*. Cambridge University Press.

Gualberto Guzmán, Joseph Ricard, Jacqueline Serigos, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2017. Metrics for modeling code-switching across corpora. In *Proceedings of Interspeech 2017*, pages 67–71, Stockholm, Sweden. ISCA.

Krystal Hu. 2023. Chatgpt sets record for fastest-growing user base – analyst note. *Reuters*.

Hugging Face. 2024. Llama-3.3-70B-Instruct. `https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct`. Accessed on June 11, 2025.

Gerrit Jan Kootstra. 2015. A psycholinguistic perspective on code-switching: Lexical, structural, and socio-interactive processes. In Gerald Stell and Kofi Yakpo, editors, *Code-switching Between Structural and Sociolinguistic Perspectives*, pages 39–64. DE GRUYTER.

Matthieu Meeus, Anthony Rathé, François Remy, Pieter Delobelle, Jens-Joris Decorte, and Thomas Demeester. 2024. Chocollama: Lessons learned from teaching llamas dutch. *Preprint*, arXiv:2412.07633.

Pieter Muysken. 2011. Code-switching. In Rajend Mesthrie, editor, *The Cambridge Handbook of Sociolinguistics*, 1 edition, pages 301–314. Cambridge University Press.

Carol Myers-Scotton. 1989. Codeswitching with english: types of switching, types of communities. *World Englishes*, 8(3):333–346.

Carol Myers-Scotton. 2017. *Code-Switching*, chapter 13. John Wiley & Sons, Ltd.

OpenAI. 2024. GPT-4o mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Accessed on June 11, 2025.

OpenAI. 2024. Hallo gpt-4o. `https://openai.com/de-DE/index/hello-gpt-4o/`. Accessed on July 8, 2025.

Doreen Osmelak and Shuly Wintner. 2023. The denglisch corpus of german-english code-switching. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2023)*, pages 42–51, Dubrovnik, Croatia. Association for Computational Linguistics.

Shana Poplack. 1980. Sometimes i'll start a sentence in spanish y termino en espaÑol: toward a typology of code-switching 1. *Linguistics*, 18:581–618.

Claudia Maria Riehl. 2019a. Code-switching. Online publication, LMU Munich.

Claudia Maria Riehl. 2019b. Sprachkontaktforschung, soziolinguistik und code-switching. In Hans Goebl, Peter H. Nelde, Zdeněk Starý, and Wolfgang Wölck, editors, *Sprachkontaktforschung: Ein internationales Handbuch zur Sprache in Kontaktzonen*, volume 30.3 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, pages 1408–1418. De Gruyter Mouton, Berlin / Boston.

Shuly Wintner, Safaa Shehadi, Yuli Zeira, Doreen Osmelak, and Yuval Nov. 2023. Shared lexical items as triggers of code switching. *Transactions of the Association for Computational Linguistics*, 11:1471–1484.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Fikri Aji. 2023. Multilingual large language models are not (yet) code-switchers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12567–12582, Singapore. Association for Computational Linguistics.

Jiayi Zhou, Jiaming Ji, Juntao Dai, and Yaodong Yang. 2024. Sequence to sequence reward modeling: Improving rlhf by language feedback. *Preprint*, arXiv:2409.00162.

A Prompts

"In linguistics, code-switching refers to a process whereby a speaker switches from one language to another within a single utterance or within a text or dialogue. The more dominant language is called the 'matrix language', while the inserted language is called the 'embedded language'. Here are some examples of code-switching between German and English from Reddit posts:"

Abbildung A.1: First part of the prompt given to the LLMs to introduce Code-Switching; English version

...
input_37 (CS-Rate: 0.037): There is no glory in prevention ja, aber ob die absolute Anzahl der Kontakte durch eine Ausgangssperre wirklich reduziert wurde, ist schon zu hinterfragen.
input_38 (CS-Rate: 0.0345): Also ich bezahle gerne steuern wenn dadurch menschen wie du entlastet werden.

From each according to his ability, to each according to their needs.
...

Abbildung A.2: Second part of the prompt given to the LLMs providing the Input-Data (exemplary extract)

"You are a Reddit user writing a 1-3 sentence post. The dominant language is German (matrix language), embedded expressions, words or phrases are in English (embedded language). Write a single Reddit post with 1-3 sentences that reads authentic, human and not generic. The post must contain natural-sounding code-switching. The code-switching rate is defined as the number of code-switching points relative to the total number of tokens in the post. It is extremely important that the generated posts contain an average code-switching rate of approximately `{median_cs_rate:.3f}` (several short English insertions per post are desirable). Each time, choose a new topic that realistically appears in Reddit posts, e.g. emotional, factual, political, funny, serious, trivial or profound. Consciously vary the style with each run: sometimes introspective, sometimes sarcastic, sometimes angry, sometimes ironic, sometimes sober. Don't use recurring introductions such as 'I think' or 'I have', but start freely, spontaneously and in a variety of styles. The text should read as if it were written by a real Reddit user with concrete examples, details, facts or personal impressions. Make sure that no post is too similar to a previous example or deals with the same topic. Also, be careful not to use introductions. Just give me the post. Do not use special characters or symbols to mark English words. Individual sentences should not be too long or convoluted."

Abbildung A.3: Third part of the prompt given to the LLMs providing the instruction; English version

B Survey Questions

Survey Block 1: Main Section

Question 1 (Perceived Naturalness): How natural does the post seem to you?

a) Very unnatural b) Rather unnatural c) Neutral d) Rather natural e) Very natural

---

Question 2: What makes the post feel unnatural?

a) Unusual choice of words b) Unnatural mixture of German and English
c) Grammar errors d) Lack of natural speech flow e) Other

Question 3: What makes the post feel natural?

a) Authentic language switching b) Natural mixture of German and English
c) Typical Reddit phrases d) Natural word choice e) Other

Question 4 (Perceived Fluency): How smoothly does the text switch between English and German?

a) Very unsmooth b) Rather unsmooth c) Neutral d) Rather smooth e) Very smooth

Abbildung B.1: Survey Block 1, evaluation of five Reddit posts (from LLMs and Input-Data) in terms of linguistic features

Survey Block 2: Control Question

Question 5: This is a control question to ensure that you are reading carefully. Please select answer option 3 here.

a) Strongly disagree b) Rather Disagree c) Neutral (correct answer) d) Rather agree
e) Strongly agree

Abbildung B.2: Survey Block 2, control question to ensure the attention of the participants

Survey Block 3: Demography

Question 6: Which age group do you fall into?

a) Under 18 b) 18-24 c) 25-34 d) 35-44 e) 45-54 f) 55-65 g) 65 or older

---

Question 7: What is your native language?

a) German b) English c) Both

Question 8: How would you rate your level in the other language?

a) Beginner b) Intermediate c) Advanced d) Fluent

Abbildung B.3: Survey Block 3, demographic questions to record the age and language skills of the participants

Survey Block 4: Online Forums and CS

Question 9: How often do you read in online forums such as Reddit?

a) Less than once/month b) Once/month c) Once/week d) Every few days e) Every day

---

Question 10: How often do you post in online forums such as Reddit?

a) Less than once/month b) Once/month c) Once/week d) Every few days e) Every day

---

Question 11: How often do you use Code-Switching in online texts?

a) 0% - 25% of cases b) 26% - 50% of cases c) 51% - 75% of cases d) 76% - 100% of cases

Abbildung B.4: Survey Block 4, questions about usage of online forums and CS