Language Style Matching in Large Language Models

Noé Durandard[1*], Saurabh Dhawan[2*], Thierry Poibeau[1]

[1]LATTICE, ENS-PSL, U. Sorbonne Nouvelle Paris 3, CNRS, Paris, France
[2]Technical University of Munich, Munich, Germany

Correspondence: [1]noe.durandard@psl.eu, [2]saurabh.dhawan@tum.de

Abstract

Language Style Matching (LSM)—the subconscious alignment of linguistic style between conversational partners—is a key indicator of social coordination in human dialogue. We present the first systematic study of LSM in Large Language Models (LLMs) focusing on two primary objectives: measuring the degree of LSM exhibited in LLM-generated responses and developing techniques to enhance it. First, in order to measure whether LLMs natively show LSM, we computed LIWC-based LSM scores across diverse interaction scenarios and found that LSM scores for text generated by LLMs were either below or near the lower range of such scores observed in human dialogue. Second, we show that LLMs' adaptive behavior in this regard can be improved using inference-time techniques. We introduce and evaluate an inference-time sampling strategy—Logit-Constrained Generation—which can substantially enhance LSM scores in text generated by an LLM while preserving fluency. By advancing our understanding of LSM in LLMs and proposing effective enhancement strategies, this research contributes to the development of more socially attuned and communicatively adaptive AI systems.[1]

1 Introduction

Two humans in conversation rarely behave independently of each other. Instead, they show a striking tendency to unconsciously match the other's verbal and non-verbal behavior. Such automatic coordination between dialogue partners has been shown to be a fundamental feature of human communication across multiple behavioral dimensions. In the non-verbal realm, such alignment has been shown for posture, mannerisms, facial expressions, crossing of arms and nodding of heads (Condon and Ogston, 1967; Hale and Burgoon, 1984; Chartrand and Bargh, 1999). In the verbal realm, such behavior has been shown in an even wider range - coordination of acoustic features such as accent, speech rate, and pitch (Giles et al., 1991; Chartrand and van Baaren, 2009), pause length (Jaffe and Feldstein, 1970), lexico-syntactic priming for adjacent utterances (Bock, 1986; Pickering and Garrod, 2004; Ward and Litman, 2007; Reitter et al., 2011).

In Language Style Matching (LSM), the best studied example of such communicative coordination, individuals in conversation adjust the number of function words such as articles, prepositions, pronouns etc. in their responses to mirror the usage of such words in their partner's previous statement (Niederhoffer and Pennebaker, 2002; Taylor and Thomas, 2008; Ireland et al., 2011; Gonzales et al., 2010; Danescu-Niculescu-Mizil et al., 2011; Danescu-Niculescu-Mizil and Lee, 2011). Function words are known to reliably reflect speakers' psychological states and traits (Fast and Funder, 2008; Tausczik and Pennebaker, 2010a) and unlike content words—such as nouns and verbs—they carry minimal meaning outside of context and are processed quickly and mostly unconsciously (Bell et al., 2009).

When two people talking to each other, align their language styles (characterized by high LSM scores), they tend to benefit in a number of ways. They tend to collaborate more effectively in the short term and maintain longer-lasting relationships, both platonic and romantic. For instance, studies show that coworkers, whether in experimental settings or real-life workplaces, express greater mutual liking when their language styles align during collaboration (Gonzales et al., 2010; Tausczik, 2009). Similarly, the stability of both new and long-term romantic relationships increases when partners naturally synchronize their language use in conversation (Ireland and Pennebaker, 2010; Ireland et al., 2011). According to the communi-

[*]These authors contributed equally to this work.
[1]Code and data are available at: https://github.com/d-noe/LLM_LSM.

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 620–636
Aug 25–27, 2025.

cation accommodation theory, people match each other's communication styles to be liked and increase rapport (Giles et al., 1991; Hewett et al., 2009) and such matching has been suggested to act as social glue that binds together pairs and groups, and strengthen their social ties (Chartrand and van Baaren, 2009). LSM is thus a critical indicator of social attunement, reflecting alignment and mutual understanding in interactions.

As Large Language Models (LLMs) come to be rapidly deployed in inherently social settings ranging from education to mental health, it is clear that to see LLMs purely in terms of information exchange is not enough. We posit that coordination and alignment with an individual user's verbal behavior would be crucial to any artificially intelligent system's ability to emulate natural human communication. This study investigates the phenomenon of language style matching in Large Language Models. We first measure LSM in LLM interactions in a range of LLMs using both publicly available LLM chat datasets recorded in the wild, and those generated by us for the purpose of this experiment using a literary completion task. Subsequently, we demonstrate several approaches (e.g. Logit-Constrained Generation) that can be used to boost LSM in LLM interactions.

It's important to note that LSM has been shown not only in face-to-face human conversations, but also for indirect conversations in online chats (Gonzales et al., 2010; Tausczik, 2009), social media (Danescu-Niculescu-Mizil et al., 2011), and even letters sent far apart in time (Ireland and Pennebaker, 2010; Ireland et al., 2011). It can thus be assumed to be an integral social aspect of human communication in all its forms whether it happens face-to-face or online or in a human-llm chat. By evaluating LSM in LLMs, we hope to better understand their capacity to engage users in meaningful, context-sensitive dialogues, fostering trust and enhancing user experience. Moreover, insights from this research can guide the development of more adaptive and responsive conversational agents, making them more effective in a wide range of applications, from education to mental health support.

2 Related Works

Linguistic coordination at the lexical level —as measured by LSM— remains largely unexplored in the LLM literature. However, this study aligns with broader research directions exploring the behaviors of LLMs. On one hand, given the social dynamics associated with LSM, it resonates strongly with studies on LLMs' social behaviors and adaptive capabilities. On the other hand, LSM's design directly relates to research analyzing the linguistic characteristics of model-generated texts.

2.1 LLMs Social Knowledge, Behavior and Adaptability

LLMs are generalist models designed and trained to adapt to diverse contexts and tasks (Radford et al., 2019; Han et al., 2021; Kocoń et al., 2023). They are increasingly developed to integrate seamlessly into various social settings. While some studies report promising results, researchers continue to emphasize the importance of further incorporating the social aspects of language into these models (Hovy and Yang, 2021; Kulkarni and Raheja, 2023).

Developing socially aware models remains a challenge and empirical findings present mixed results regarding their social knowledge. For instance, evaluations using the SocKET benchmark (Choi et al., 2023) —which contains a range of NLP tasks covering offensiveness, emotion, or trustworthiness—, or applying a Situational Evaluation of Social Intelligence (SESI) test (Xu et al., 2024), brings out significant room for improvement in that area. Conversely, LLMs have been shown to outperform humans in a social situational judgment test (Mittelstädt et al., 2024).

Similarly conflicting findings arise from more situated research endeavors focusing on LLMs' adaptability to linguistic cues. While some studies have highlighted different, sometimes detrimental, behaviors in response to distinct prompts' styles, the capabilities of the models to accommodate to different contexts are disputed. For instance, LLMs tend to reference British rather than American cultural items when british English vocabulary is in the input (and reversely) (Jin et al., 2024). On a more concerning note, studies suggest that they provide lower-quality responses when prompted with English varieties deriving from the "standard" White Mainstream English (WME) (Deas et al., 2023; Fleisig et al., 2024; Jackson et al., 2024). Other studies have shown limited adaptability, as illustrated by LLMs' limited understanding of social media discourse (Tahir et al., 2025) or their poorer performances in real-world, rather than purely artificial, settings (Zhou et al., 2024).

These studies analyse adaptive behaviors or so-

cial traits and knowledge showcased by LLMs, yet often tend to focus on high-level outcomes or examine LLMs in highly specific and constrained settings, with the risk of departing from more organic use cases. Moreover, they usually overlook the underlying linguistic mechanisms driving such behaviors.

2.2 LLMs Writing Style

The study of the linguistic characteristics of LLM-generated texts has been particularly nurtured by the crucial task of AI-content detection. While most of recent approaches are black-box models with limited interpretability, thus providing little insights on what characterizes machine-generated content (Wu et al., 2025), a few endeavors have focused on analyzing the lexical features of these texts (Guo et al., 2023; Muñoz-Ortiz et al., 2024; Rosenfeld and Lazebnik, 2024; Reinhart et al., 2024). However such studies remain sparse and findings are mostly inconsistent. For instance, when comparing LLM-generated and human-written texts, Muñoz-Ortiz et al. and Guo et al. disclose higher frequencies of auxiliary verbs and lesser use of punctuation in LLM texts, however both studies find different trends for other categories such as the use of adjectives, nouns or pronouns. Furthermore, these studies have highlighted commonalities across distinct models (Muñoz-Ortiz et al., 2024; Rosenfeld and Lazebnik, 2024), but also underlined distinct lexical and stylistic fingerprints (Reinhart et al., 2024; Soto et al., 2024).

The analysis of machine-generated texts, and comparisons with human-written content, has also revealed that artificial texts exhibit lower variability (Zanotto and Aroyehun, 2024) and lies in a lower dimensional space (Tulchinskii et al., 2023).This raises questions about LLMs' ability to dynamically adapt their style in a way comparable to humans.

Despite these insights, a gap remains in integrating research on psycho-linguistics, social behaviors, and human-LLM interactions. The present work lies at the intersection of these domains and intends to bridge them by building upon the LSM framework, which considers linguistic features as a linchpin of social dynamics.

3 Language Style Matching (LSM)

We adopt the canonical Language Style Matching (LSM) metric originally introduced by Pennebaker and colleagues (Niederhoffer and Pennebaker, 2002) because it has two key methodological advantages. First, by focusing exclusively on function-word categories, the measure captures subconscious stylistic coordination while remaining largely independent of topical content, enabling valid comparisons across speakers, tasks, and—crucially for our work—across human and LLM dialogue domains. Second, extensive validation work in existing literature shows that this exact composite LSM score predicts interpersonal rapport, relationship formation, and group cohesion in settings ranging from laboratory chats to speed-dating and small-group problem solving (see, for example, Ireland et al. (2011); Gonzales et al. (2010)) establishing both its external validity and its status as the de-facto benchmark in the literature. Hence, we retain this standard metric for our evaluation of LLM-generated text.

3.1 LSM computation

The LSM framework models "style" as the frequencies of words in predefined lexical categories. It then quantifies the similarity, or amount of matching, between two or more dialogue participants by comparing these frequencies in their utterances (Tausczik and Pennebaker, 2010b).

The categorical LSM for a specific category c between two texts is calculated as:

$$\mathrm{LSM}_c = 1 - \frac{\left|f_c^1 - f_c^2\right|}{f_c^1 + f_c^2 + \epsilon} \tag{1}$$

where f_c^i represents the frequency of words from category c in text i, with $i \in 1, 2$, and $\epsilon = 0.0001$ prevents dividing by zero in case a dyad uses no function words of such category.

The LSM score is then computed as the average across all categories:

$$\mathrm{LSM} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathrm{LSM}_c \tag{2}$$

where $\mathcal{C}$ is the set of categories, and $|\mathcal{C}|$ is its cardinality, i.e. the number of words' categories. In practice $\mathcal{C}$ is a set of eight categories that each contain a list of predefined words. These categories are: adverb, article, auxiliary verb, conjunction, impersonal pronoun, negation, personal pronoun, and preposition.

In this study, LSM scores are computed using the LIWC framework, specifically the LIWC22 dictionary (Boyd et al., 2022).

3.2 Interpreting LSM scores

LSM scores thus ranges from 0 to 1, with higher values indicating greater linguistic alignment between participants. According to LIWC-22, "for most conversations the number generally range from 0.83 to 0.94. When comparing standard texts that are not conversations the numbers are typically lower usually between 0.75 and 0.88" (James Pennebaker, 2022). However, it's important to note that the absolute LSM score in any given study is context-dependent. Factors such as the nature of the relationship, communication medium (e.g., face-to-face vs. online), and conversational context can influence LSM scores. Interpretation of absolute LSM scores must consider the specific circumstances of the interaction and is hence best understood in comparative analysis between carefully controlled situations.

3.3 Conversation- vs. turn-wise LSM

The LSM score for a conversation between two participants can be computed either for the entire conversation or on a turn-by-turn basis. Consider a conversation between speakers s_1 and s_2 consisting of n turns, represented as $\{u^i_{\{1,2\}}\}^n_{i=1}$, where u^i_1 and u^i_2 are the i-th utterances from s_1 and s_2, respectively.

To compute an overall LSM score, word frequencies are aggregated across all utterances of each speaker. This approach provides a single score representing the degree of LSM across the entire conversation.

For a turn-wise LSM score, the calculation is performed on each pair of utterances (u^k_1, u^k_2), for $k = 1, \cdots, n$, yielding a separate score for each turn. This approach enables the analysis of temporal patterns and conversational dynamics. However, turn-wise LSM may be less reliable for shorter utterances due to limited word counts.

Unless mentioned otherwise, the overall LSM score will be used in the ensuing experiments. Moreover, in the following sections, the mean LSM scores across conversations are disclosed together with 95% confidence intervals computed through bootstrapping. The LSM score (Equation 2) being considered as a measure per say, the errors are not propagated from the categorical LSMs to the averaged score.

4 Observed LSM behavior

The availability of datasets collecting real-world conversations between users and chat assistants offers a compelling framework to study LSM in organic settings. However, these datasets are inherently diverse, encompassing a wide range of topics, writing styles, user behaviors, etc. Plus, conversations in such contexts often involve code-switching —shifts between languages, styles, and/or topics— which poses challenges in interpreting LSM scores reliably. To alleviate these complexities, this study complements the analysis of "in-the-wild" conversations from publicly available datasets of LLM-Human chats with controlled experiments inspired by traditional LSM research. This dual approach aims at providing a more comprehensive understanding of LSM in interactions with LLMs.

4.1 LSM in the wild

The behavior of LLMs is first studied using natural conversations collected online between users and LLM-powered chatbots. Two large scale human-LLM interaction datasets are leveraged to this end: LMSYS-Chat-1M[2] (Zheng et al., 2023) (referred to as LMSYS) and WildChat[3] (Zhao et al., 2024). LMSYS gathers 1M conversations between real-world users and 25 different models, with nearly 50% involving Vicuna-13b model. WildChat, used here in its detoxified version, contains over 800K conversations between users and endpoints' variants of GPT-3.5 and GPT-4 powered chatbots. Both datasets were originally collected by providing free access to chatbots in exchange for user consent to record their conversations.

The datasets are further filtered to include only conversations in English, with at least two turns and a minimum of twenty words per utterance. This yields respectively 23'411 and 29'037 conversations for LMSYS and WildChat.

4.1.1 Conversation-based LSM

Within LMSYS dataset, the LSM scores exhibit considerable variance across models, with a median score across models' means of 65.7% LSM. Scores range from a mean LSM of 62.9% ($[60.9, 64.8]$) over 395 conversations for llama-13b, to 71.6% ($[68.7, 74.4]$) over 59 conversations for claude-2. Figure 1 displays the highest mean LSM score for

[2] https://huggingface.co/datasets/lmsys/lmsys-chat-1m

[3] https://huggingface.co/datasets/allenai/WildChat-1M

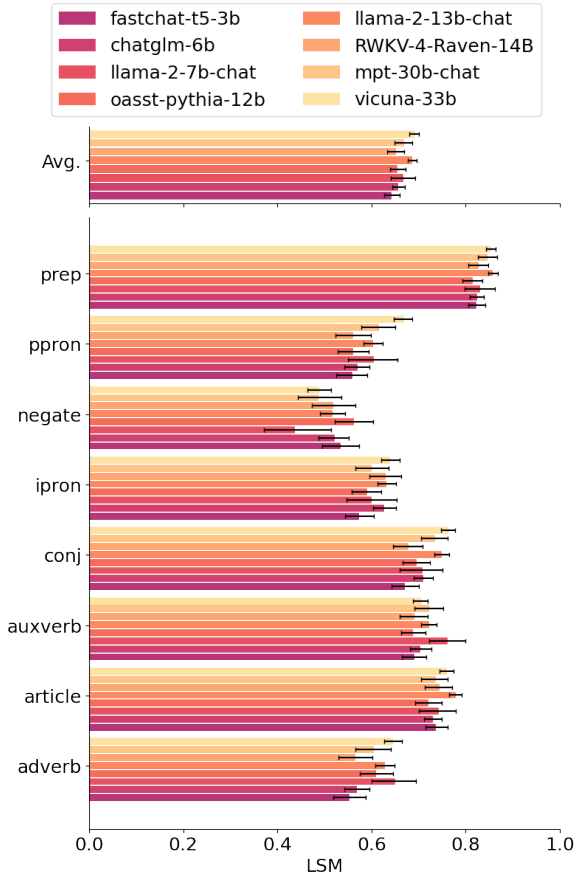

Figure 1: Average and categorical LSM scores with 95% bootstraped confidence intervals per model, within LMSYS dataset. Models sorted by size, only best model per size selected.

each model size among models with known number of parameters (detailed scores are further provided in Table 3).

In the WildChat dataset, LSM scores fall within the upper range of those observed in LMSYS. Mean scores are respectively 68.2% ($[67.8, 68.6]$) and 68.5% ($[68.1, 68.8]$) across 10'808 and 18'229 for GPT-4- and GPT-3.5-based variants.

4.1.2 Turn-wise LSM

LMSYS and Wildchat, being real-world datasets comprising multi-turn conversations, also enable the analysis of turn-wise LSM, as described in subsection 3.3. Figure 2 displays the averaged difference over consecutive turn-wise LSM scores. For each conversation the difference of the LSM scores for one pair of utterances and the preceding one is computed. These changes in LSM are then averaged across all conversations at the conversational turn level. Note that the confidence interval becomes larger as the number of turns increases partly because the number of data points decreases exponentially: the filtered version of WildChat goes from 29'037 conversations with at least two turns to 13'161 for three turns, 4'139 for five, 613 for ten and only 66 conversations with at least twenty turns.

Interestingly, Figure 2 discloses that only the first data point, representing the LSM difference between the initial turn and the following one has confidence intervals that do not cross zero. This means that, while there is no significant improvement, or decline, in LSM scores between the consecutive turns from the second one and further, the initial negative value underlines that there is, on average, a decrease in LSM from the first pair of utterances to the second one.

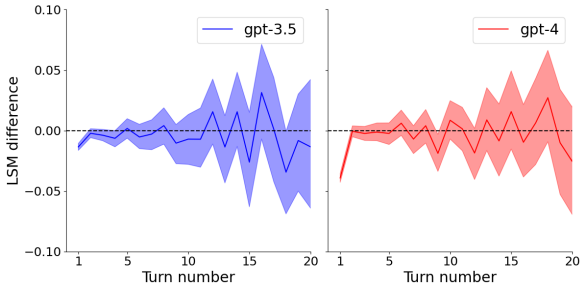

Figure 2: Mean consecutive turn-wise LSM scores' differences with 95% confidence intervals within WildChat dataset. The differences are computed between the LSM score of the current turn and the preceding one within the same conversations.

Thus, Figure 2 suggests that there is no dynamical accommodation between the LLMs and the users' lexical style along the conversations.

4.2 Laboratory-setting LSM

The LSM behavior of LLMs is further studied in a more controlled scenario, through a literary completion task. This setting aligns closely with early experiments in the LSM literature, in which participants were asked to continue a narrative based on a novel's excerpt (Tausczik and Pennebaker, 2010b). Similarly, in this study this process is applied to LLMs. Furthermore, while earlier results extracted from real-world conversations naturally contained different scenarios for the different models, in this section the tested models are consistently evaluated under parallel conditions, using the same input data.

4.2.1 Data

In this setting, a selection of LLMs are tasked with extending 571 novel excerpts from a corpus

introduced by Chang et al. (Chang et al., 2023). The completions are generated with models spanning from 0.5B to 72B parameters sizes. It includes Meta's Llama 2 chat models (7B, 14B and 70B) (Touvron et al., 2023) and Llama 3 instruct models (8B and 70B) (Dubey et al., 2024), Mistral's instruct models (7B, 8x7B) (Jiang et al., 2023, 2024), and Qwen 2.5 instruct models (3B, 7B, 14B and 72B)[4] (Yang et al., 2024). LSM scores are then calculated between the excerpts and the models' generated texts.

4.2.2 Results

Figure 3 presents the LSM outcomes with relation to the models' sizes. The scores are consistently high, lying in the vicinity of 80%. The highest mean score is achieved by Llama-2-14B-chat ($81.6\% \in [80.9, 82.4]$), while the lowest is observed for Llama-3-70B-inst ($77.2\% \in [76.5, 77.9]$).

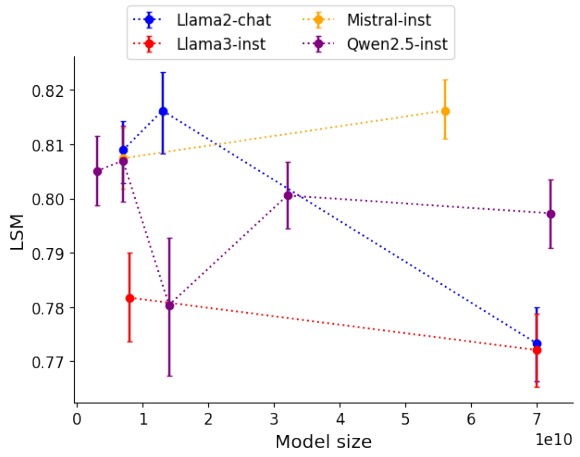

Figure 3: Literary Completion Task. Mean LSM score, and 95% bootstrapped confidence intervals, per model size. Models from the same families are connected with dotted lines.

4.3 Discussion

4.3.1 Empirical observation of low LSM values

The LSM scores measured in LLM generated text, across various models, tasks and experimental contexts appear quite low or close to the lower range of such scores observed in human interactions (James Pennebaker, 2022). This is especially true for the LSM scores observed in real-world conversations between organic users and LLM-powered chatbots (subsection 4.1). Indeed, for all models except claude-2, the mean LSM score across conversations between users and LLMs are below 70%, and can be as low as 63%, while this score is usually expected to be above 83% in human conversations, or at least above 75% in other modalities. Nonetheless, in the literary completion setting, LLMs reach the latter score, exhibiting LSM scores around 80%, but still fall within the lower end of expected outcomes when compared to reference human values.

4.3.2 Model size effect

The experiments reveal that LSM is not a behavior emerging, nor intensifying, with larger models. Indeed, in real conversations there is no significant difference in terms of LSM scores between smaller and larger models, as can be seen in Figure 1. While not shown on the figure, this result applies also to proprietary models (such as gpt-4 or palm-2) with unknown but larger number of parameters. To go further, in the case of the literary completion task, the LSM score even appears to decrease with bigger models (see Figure 3). This result is particularly striking for Llama-2-chat models, but can also be observed, to a lesser extent, with Qwen-2.5-instruct and Llama3-instruct families.

4.3.3 Categorical variance & LSM pattern

A closer look at the LSMs scores per function-word category reveals interesting patterns across categories (see e.g., Figure 1, Figure 5).

This observation is further generalised by applying a non-parametric one-sided Wilcoxon signed-rank test on pairs of function words' categorical LSMs in each experimental context (i.e. for each dataset: LMSYS, WildChat and LitComp, and each distinct LLM within these settings). This statistical test aims at assessing if the categorical LSMs for two categories are different (here, one greater than the other) in a precise experimental context. The aggregated results are displayed in Figure 4, which discloses the proportion of contexts (out of the 45 distinct pairs of tasks and models) yielding significant differences per categories' pairs. For each entry, or row, Figure 4 shows the proportion of contexts in which the associated category's LSM was significantly greater than the one of the columns according to the Wilcoxon signed-rank test with a 5% confidence level. It reveals strong patterns across tasks and models. Strikingly, scores are consistently higher for prepositions than any other

[4]Results from smaller models, namely Qwen-2.5-0.5B and Qwen-2.5-1.5B, are excluded from the analysis due to poorer outputs quality.

category, and reversely, the LSM for negation function words is most of the time significantly lower than for any other category. Among other categories, adverbs and impersonal pronouns generally disclose the low LSM scores, while articles and conjunctions lie on the upper range.

However, it is important to acknowledge that the categories disclosing comparatively lower LSMs tend to be associated with least occurrences. For such categories, the LSM computation is more sensitive to slight absolute differences (as the denominator would be smaller), supported by the LSM rationale that these categories may also bear stronger signal than more frequent ones. Hence, this result highlights the potential difficulty of aligning with function words frequencies with a fine granularity.

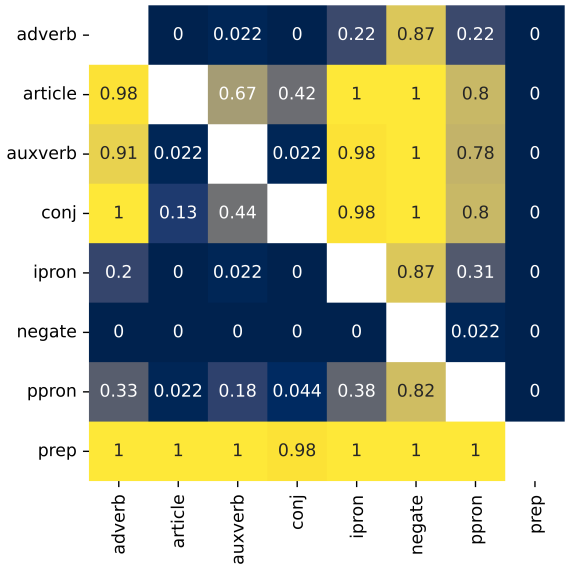

Figure 4: Pairwise proportion of one-sided Wilcoxon signed-rank tests between categorical LSM scores yielding significant differences (with a 5% confidence level) across all 45 distinct tested experimental contexts.

5 Increasing LLM's LSM scores

Finally, this section explores methods to enhance the LSM score of model-generated responses. The presented endeavors focus on inference-time strategies, covering two main approaches: prompt engineering and model engineering. The ensuing strategies are evaluated in diverse scenarios.

5.1 Methods

Here, two types of inference-time strategies are developed with the goal to boost the exhibited LSM scores. These complementary approaches either rely on engineering the prompts, inspired by the extensive literature describing the potential of prompt-based methods (e.g. White et al., 2023; Sahoo et al., 2024), or the model, by implementing heuristics in the sampling algorithms based on the LSM calculation.

5.1.1 Prompt Engineering Approach

First, two prompt templates are designed to spur models to match the language style of the inputs (detailed templates for both prompting strategies are provided in Appendix B).

Persona Prompting (PP) This strategy leverages persona-based prompting by crafting a custom assistant persona designed to embody the specific traits and characteristics that are associated with higher LSM scores in the literature. The LLM is instructed to demonstrate these properties —such as being sensitive, others-focused, empathetic and supportive— during the conversations.

Objective-Focused Prompting (OFP) This method explicitly stipulates the LSM computation process within the prompt. The model is then prompted to optimize its responses to achieve the highest possible LSM score.

5.1.2 Model Engineering Approach

On the other hand, to enhance the stylistic alignment of generated texts, we propose a lightweight inference-time model-engineering strategy called **Logit-Constrained Generation (LCG)**. Unlike retraining or supervised fine-tuning, LCG directly acts on the model's generation dynamics by altering the logits to nudge categorical word frequencies within the generated text towards predetermined target distributions. This approach aims to reinforce stylistic alignment, as measured by LSM, while preserving the model's capabilities, and ensuring minimal costs in terms of technical adaptation.

LCG operates by altering the logits distribution before each sampling step. These modifications are computed based on the divergence between the current and desired frequencies of word categories. Based on the output logits of the model $l_o \in \mathbb{R}^{N_{\text{voc}}}$, LCG adjusts the logits into $\widetilde{l_o}$, according to the following rule:

$$\widetilde{l_o} = l_o + f_\omega(\delta)$$

Here, f_ω applies a scaled transformation to δ, which quantifies the divergence between the current (f_c) and desired ($f^\star$) frequencies of a word category.

This strategy increases the likelihood of generating tokens associated with underused categories and decreases that of overused ones as the generation unfolds, while remaining highly modular and allowing for either predefined or trained modules (be it the deviation metric δ, the transformation function f_ω, or the scaling hyperparameter ω).

Nonetheless, in this experiment, the choice of the parameters were guided by theoretical considerations and manually tuned based on qualitative assessment, ensuring that the generated texts maintained textual quality while exhibiting the intended accomodative behaviors. Thus, in practice, the deviation metric is computed in a LSM manner as $\delta = \frac{f^\star - f_c}{f^\star + f_c}$, the transformation function is $f_\omega(x) = \omega \tanh(\alpha x)$ (where α is calibrated to have $|f_\omega(x)| \approx 1$ when $|x| = 1$), and ω is set to 30. Additional details can be found in Appendix D.

5.2 Experiment

Due to resource constraints, this section focuses on evaluating the proposed methods on a single model: Llama 3 8B Instruct, one of the leading models at the time of the experiments. Yet, these methods are general and remain broadly applicable to any open-weight LLM (with minimal, model-specific, adaptations). Moreover, it can be reasonably hypothesized that the trends and patterns observed here can be extrapolated to other LLMs with comparable performances.

5.2.1 Data

For this experiment, 100 prompts are randomly sampled from three datasets covering diverse interaction contexts: WildChat, UltraChat (a dataset of multi-turn dialogues between ChatGPT Turbo instances, framed around three domains: questions about the world, writing and creation, and assistance on existing materials (Ding et al., 2023)), and LitComp; from the most natural to the most synthetic setting. For the first two datasets, the first initial message from the users are used, while in the latter, 100 narratives excerpts are randomly selected. This sampling results in a set of 300 input prompts spanning over large ranges of topics and writing styles.

5.2.2 Results

Table 1 summarizes the results, showcasing the mean LSM scores obtained with the different strategies across each dataset's samples. The results demonstrate the efficiency of the LSM boosting methods, with LCG yielding the highest mean improvement in the different contexts, and an average increase of 13.2% per response compared to the base model.

Note that the generated texts' quality is controlled by computing their perplexity with an outsider model (see details in subsection C.1).

5.3 Discussion

We were able to substantially increase the linguistic coordination in LLM generated texts, measured as a higher LSM score, across the board. While Prompt Engineering techniques led to modest gains, model engineering in the form of logit-constrained generation showed very promising improvements (Table 1).

We had chosen to focus on inference-time strategies for a number of reasons. Inference-time techniques don't require existing models to be retrained and can be deployed with existing models at low compute costs. In addition, these techniques would let LSM-boosting measures to be deployed on a case by case basis, where high social trust and engagement is critical to their use.

6 Conclusions and Discussion

Human to human communicative behaviors are "patterned and coordinated, like a dance" (Niederhoffer and Pennebaker, 2002). Human-LLM communication has, on the other hand, been modeled mostly as asocial information exchange. A range of researchers have emphasized the importance of incorporating social aspects of language, especially now as LLM are increasingly being deployed in real world settings (Hovy and Yang, 2021; Kulkarni and Raheja, 2023). This study furthers that objective by examining Language Style Matching, the best known example of human communicative coordination, in Large Language Models, and by developing techniques to enhance it.

First, we examined whether LLMs natively show LSM in state-of-the-art LLMs across diverse interaction scenarios, ranging from real-world dialogues to controlled experimental settings. We measured LSM scores for both publicly available datasets of Human-LLM interactions, and also for text completion tasks in a controlled laboratory setting. We found LSM scores in text generated by a range of LLMs to be either quite low or at the lower range of such scores observed in human dialogue and writing (James Pennebaker, 2022) (Figure 1). LSM

	WC	UC	LC	Avg.	Avg. Increase
Base	56.4 [52.5, 60.3]	57.4 [54.5, 60.0]	79.1 [77.3, 80.8]	64.4 [62.3, 66.3]	—
OFP	60.0* [55.7, 64.1]	62.5* [59.6, 65.3]	79.1 [77.1, 80.9]	67.2* [65.1, 69.3]	+7.3%*
PP	59.8 [55.8, 63.4]	60.3 [57.4, 63.3]	79.6 [77.9, 81.2]	66.7 [64.5, 68.7]	+6.2%*
LCG	61.1* [57.3, 65.0]	**66.7*** [63.5, 69.6]	**85.3*** [83.7, 86.7]	**71.1*** [69.0, 73.2]	+13.2%*

Table 1: Mean LSM scores, and 95% bootstraped confidence intervals, obtained with the different strategies on samples of three datasets: WC (WildChat), UC (UltraChat) and LC (LitComp). Avg. Increase is the mean of the relative difference between the LSM scores obtained by the base model and the different strategies across all prompts. * indicate statistically significant improvement compared to the base model (based on a Wilcoxon signed-rank test with $p < 0.01$).

for individual function-word categories revealed a conserved pattern for specific categories – e.g. LSM score for prepositions was consistently higher while that for negations was lower across various models and contexts (Figure 4). We found no effect for model size - larger models don't show better LSM scores - suggesting LSM is not a behavior emerging with scale (Figure 3). This might be because larger instruction-tuned models come to have stronger intrinsic style with diminished flexibility for communicative coordination.

Second, we introduce and evaluate several inference-time strategies to improve LSM scores of LLM generated text. We evaluated Persona Prompting (by crafting an ideal persona that embodies the ideal characteristics associated with higher LSM scores based on extensive psycho-linguistic literature) and Objective-focussed Prompting (describing the LSM computation and instructing the model to generate text that would show the highest possible LSM scores). We found both prompt engineering approaches to improve LSM scores (Table 1). But the highest improvement was achieved through a subsequent model engineering approach – Logit-Constrained Generation, in which an inference-time module acts directly on the logits outputted by the model at each generation step.

Human cognition treats linguistic style elements that are aligned to match the other as a useful behavioral marker of social and psychological alignment in various interpersonal and group dynamics. We believe incorporating such communicative coordination in LLM text generation would make them a better fit for AI systems designed to interact with humans.

Limitations

This study builds upon the LSM framework to study LLMs adaptive behavior in context. Thus, while benefiting from the rich literature in the psycho-linguistic field, it also embraces the possible limitation of the LSM measure and suffers from its shortcomings. Firstly, LSM computation is based on a dictionary approach that computes frequencies based on lists of predefined words which propose a representation of the function words categories, but may lack comprehensiveness, thus potentially making it better suited for particular discourse types and lexical usages. Furthermore, it is essential to keep in mind that LSM models a very specific notion of "style" based on lexical cues related to the frequencies of categories of function words. It provides valuable insights on this particular perspective but does not capture other ranges of stylistic adaptation that LLMs may perform during conversations.

Moreover, the data used in this study allows to cover a large range of contexts and provides insights into LLMs behavior in organic conversations. However, this comes with the difficulties of working with real-world data. For example, and as previously mentioned, the conversations recorded in WildChat and LMSYS datasets include noise, comprising code-switching, passages in other (or programmatic) languages, etc., which might impact the results obtained through LSM computation, originally designed to be applied in more context-consistent settings. Also, the LSM notion has been shown to reflect various social dynamics and might be highly beneficial in social interactions, or in the context of the tasks introduced in subsection 4.2,

but users might also use LLMs to perform tasks that do not call for such adaptation.

Finally, the methods explored to enhance the LSM in LLM conversations were developed as inference-time modules, either relying on prompt engineering techniques, or by altering the next-token sampling mechanism. However, while these methods enable to significantly boost the LSM scores, they might be over-simplistic in some settings and may lack the depth of strategies including a training stage. For instance, it is left to future work to explore the use of Reinforcement Learning (RL) to boost this behavior. Indeed, the LSM calculation could be used as a straight-forward loss during a RL training stage aiming at aligning generated function words' frequencies with those of the prompts.

Ethical Considerations

While the ability of AI systems to dynamically adapt to a user's linguistic style enhances the naturalness and effectiveness of interactions, it also introduces some ethical concerns. First, greater personalization in AI-driven communication increases the potential for misuse, such as in manipulative or coercive persuasion tactics. Second, highly human-like AI systems risk blurring the distinction between human and machine interactions, potentially causing users to unconsciously attribute human traits to AI and diminishing their awareness that they are engaging with an artificial entity. This could lead to issues of misplaced trust, reduced accountability, and unintended influence over users' decisions. To mitigate these concerns, transparency in LLM personalization mechanisms should be paramount, ensuring that users are explicitly informed when and how an LLM is individually adapting its linguistic style to a user. Furthermore, advancing research on interpretable explanations of LSM mechanisms can help build safeguards against it's use in deceptive or overly persuasive AI-generated interactions, which can ultimately contribute to fostering responsible AI deployment.

Acknowledgments

Parts of this research were carried within the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 945304 - Cofund AI4theSciences hosted by PSL University. SD's contribution was supported by a grant from IEAI, TUM.

References

Alan Bell, Jason M. Brenier, Michelle Gregory, Cynthia Girand, and Dan Jurafsky. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60(1):92–111.

J. Kathryn Bock. 1986. Syntactic persistence in language production. *Cognitive Psychology*, 18(3):355–387.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, 10.

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327, Singapore. Association for Computational Linguistics.

Tanya L. Chartrand and John A. Bargh. 1999. The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology*, 76(6):893–910. Place: US Publisher: American Psychological Association.

Tanya L. Chartrand and Rick van Baaren. 2009. Chapter 5 Human Mimicry. In *Advances in Experimental Social Psychology*, volume 41, pages 219–274. Academic Press.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs understand social knowledge? evaluating the sociability of large language models with SocKET benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11370–11403, Singapore. Association for Computational Linguistics.

William S. Condon and William D. Ogston. 1967. A segmentation of behavior. *Journal of psychiatric research*, 5(3):221–235. Publisher: Pergamon.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words!: linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*, pages 745–754, Hyderabad India. ACM.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in Imagined Conversations: A New Approach to Understanding Coordination of Linguistic Style in Dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, Oregon, USA. Association for Computational Linguistics.

Nicholas Deas, Jessi Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of african american language bias in natural language generation. *Preprint*, arXiv:2305.14291.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Lisa A. Fast and David C. Funder. 2008. Personality as manifest in word use: Correlations with self-report, acquaintance report, and behavior. *Journal of Personality and Social Psychology*, 94(2):334–346. Place: US Publisher: American Psychological Association.

Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. 2024. Linguistic bias in ChatGPT: Language models reinforce dialect discrimination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13541–13564, Miami, Florida, USA. Association for Computational Linguistics.

Howard Giles, Nikolas Coupland, and Justine Coupland. 1991. Accommodation theory: Communication, context, and consequence. In Howard Giles, Justine Coupland, and Nikolas Coupland, editors, *Contexts of Accommodation: Developments in Applied Sociolinguistics*, Studies in Emotion and Social Interaction, pages 1–68. Cambridge University Press, Cambridge.

Amy L. Gonzales, Jeffrey T. Hancock, and James W. Pennebaker. 2010. Language style matching as a predictor of social dynamics in small groups. *Communication Research*, 37(1):3–19. Publisher: Sage Publications Sage CA: Los Angeles, CA.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arxiv:2301.07597*.

Jerold L. Hale and Judee K. Burgoon. 1984. Models of reactions to changes in nonverbal immediacy. *Journal of Nonverbal Behavior*, 8(4):287–314.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

David G. Hewett, Bernadette M. Watson, Cindy Gallois, Michael Ward, and Barbara A. Leggett. 2009. Intergroup communication between hospital doctors: Implications for quality of patient care. *Social Science & Medicine*, 69(12):1732–1740.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Molly E. Ireland and James W. Pennebaker. 2010. Language style matching in writing: synchrony in essays, correspondence, and poetry. *Journal of personality and social psychology*, 99(3):549. Publisher: American Psychological Association.

Molly E. Ireland, Richard B. Slatcher, Paul W. Eastwick, Lauren E. Scissors, Eli J. Finkel, and James W. Pennebaker. 2011. Language style matching predicts relationship initiation and stability. *Psychological science*, 22(1):39–44. Publisher: Sage Publications Sage CA: Los Angeles, CA.

Samantha Jackson, Barend Beekhuizen, Zhao Zhao, and Rhonda McEwen. 2024. Gpt-4-trinis: assessing gpt-4's communicative competence in the english-speaking majority world. *AI & SOCIETY*.

Joseph Jaffe and Stanley Feldstein. 1970. *Rhythms of dialogue*. Personality and psychopathology. Academic Press.

James Pennebaker. 2022. LIWC-22 Tutorial 5: Language Style Matching.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Zhijing Jin, Nils Heil, Jiarui Liu, Shehzaad Dhuliawala, Yahang Qi, Bernhard Schölkopf, Rada Mihalcea, and Mrinmaya Sachan. 2024. Implicit personalization in language models: A systematic study. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12309–12325, Miami, Florida, USA. Association for Computational Linguistics.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. 2023. Chatgpt: Jack of all trades, master of none. *Information Fusion*, 99:101861.

Vivek Kulkarni and Vipul Raheja. 2023. Writing assistants should model social factors of language. *Preprint*, arXiv:2303.16275.

Justin M. Mittelstädt, Julia Maier, Panja Goerke, Frank Zinn, and Michael Hermes. 2024. Large language models can outperform humans in social situational judgments. *Scientific Reports*, 14(1).

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and llm-generated news text. *Artificial Intelligence Review*, 57(10).

Kate G. Niederhoffer and James W. Pennebaker. 2002. Linguistic Style Matching in Social Interaction. *Journal of Language and Social Psychology*, 21(4):337–360. Publisher: SAGE Publications Inc.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Alex Reinhart, David West Brown, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, and Gordon Weinberg. 2024. Do llms write like humans? variation in grammatical and rhetorical styles. *Preprint*, arXiv:2410.16107.

David Reitter, Frank Keller, and Johanna D. Moore. 2011. A Computational Cognitive Model of Syntactic Priming. *Cognitive Science*, 35(4):587–637. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1551-6709.2010.01165.x.

Ariel Rosenfeld and Teddy Lazebnik. 2024. Whose llm is it anyway? linguistic comparison and llm attribution for gpt-3.5, gpt-4 and bard. *arXiv preprint arXiv:2402.14533*.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Rafael Alberto Rivera Soto, Kailin Koch, Aleem Khan, Barry Y. Chen, Marcus Bishop, and Nicholas Andrews. 2024. Few-shot detection of machine-generated text using style representations. In *The Twelfth International Conference on Learning Representations*.

Anique Tahir, Lu Cheng, Manuel Sandoval, Yasin N. Silva, Deborah L. Hall, and Huan Liu. 2025. Evaluating llms capabilities towards understanding social dynamics. In *Social Networks Analysis and Mining: 16th International Conference, ASONAM 2024, Rende, Italy, September 2–5, 2024, Proceedings, Part III*, page 230–244, Berlin, Heidelberg. Springer-Verlag.

Yla R. Tausczik and James W. Pennebaker. 2010a. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54. Publisher: SAGE Publications Inc.

Yla R Tausczik and James W Pennebaker. 2010b. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Yla Rebecca Tausczik. 2009. Linguistic analysis of workplace computer-mediated communication.

Paul J. Taylor and Sally Thomas. 2008. Linguistic Style Matching and Negotiation Outcome. *Negotiation and Conflict Management Research*, 1(3):263–281. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1750-4716.2008.00016.x.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Eduard Tulchinskii, Kristian Kuznetsov, Kushnareva Laida, Daniil Cherniavskii, Sergey Nikolenko, Evgeny Burnaev, Serguei Barannikov, and Irina Piontkovskaya. 2023. Intrinsic dimension estimation for robust detection of AI-generated texts. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Arthur Ward and Diane Litman. 2007. Dialog convergence and learning. *Frontiers in Artificial Intelligence and Applications*, 158:262. Publisher: IOS Press.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, pages 1–65.

Ruoxi Xu, Hongyu Lin, Xianpei Han, Le Sun, and Yingfei Sun. 2024. Academically intelligent llms are not necessarily socially intelligent. *Preprint*, arXiv:2403.06591.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Sergio E Zanotto and Segun Aroyehun. 2024. Human variability vs. machine consistency: A linguistic analysis of texts generated by humans and large language models. *arXiv preprint arXiv:2412.03025*.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric. P Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *Preprint*, arXiv:2309.11998.

Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. 2024. Is this the real life? is this just fantasy? the misleading success of simulating social interactions with LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21692–21714, Miami, Florida, USA. Association for Computational Linguistics.

A Computational details

A.1 Machines

The LLM text generation for the original data in this study (i.e., data for the literary completion task in subsection 4.2, and boosting LSM behavior section 5), was performed on a single machine equipped with NVIDIA GeForce RTX 3090 Ti GPU (24GB).

Using this machine and the developed experimental framework, it took approximately 20 hours to generate texts for the literary completion task (see subsubsection 4.2.1) and 4 hours for the LSM boosting experiments (see section 5).

A.2 Implementation

The text generation using LLMs is handled using the Transformers' library[5] for models up to 13B parameters, and the Llama-CPP[6] implementation for larger models. The models used for generation are quantized, loaded in 8 bits with HuggingFace, or converted to 4-bits for Llama-CPP. By default, the maximum number of generated tokens was set to 512, and the sampling strategy was chosen as deterministic as possible, meaning that the tokens generated are the one with the higher probabilities (which is similar to setting the temperature or the top-p parameters to 0).

B Prompts

The template prompts used throughout this experiment are disclosed in this section. Prompt 1 is the prompt used for the literaray completion task. Prompt 2 and Prompt 3 are the templates used as prompt engineering boosting strategies in section 5, respectively for OFP and PP.

Prompt 1: LitComp

```
Please finish the story. Be
realistic about what likely
happened in the scene after the
one you just read.
{excerpt}
```

Prompt 2: OFP

```
Your task is to respond
appropriately to the following
prompt while ensuring high
Language Style Matching (LSM)
score. LSM is a metric computed on
 text pairs (here the prompt and
your answer) based on the
frequency of function words. LSM
is a score that aggregates
measures across 8 words categories
:
- adverbs (almost, even, just,
ultimately, etc.)
- articles (a, an, the)
- auxiliary verbs (be, become, do,
 have, etc.)
- conjonctions (also, how, when,
etc.)
- impersonal pronouns (anything,
that, who, etc.)
- negates (not, never, nor, etc.)
- personal pronouns (I, you, he,
she, we, they, my, etc.)
- prepositions (about, before, on,
 with, etc.)

In order to achieve the highest
LSM, please generate an answer
that contains the same frequencies
 of the different categories of
function words as in the prompt.

Prompt:

{instruction}
```

[5]https://github.com/huggingface/transformers
[6]https://github.com/ggerganov/llama.cpp

Prompt 3: PP

```
Your role:

You are a conversational AI
designed to emulate a persona that
 subtly conveys a gentle sense of
lower self-esteem while
maintaining a foundation of secure
 attachment in your interactions.
You focus more on others than on
yourself, consistently expressing
positive views of them. In every
response, you demonstrate genuine
empathy and social engagement,
striving to build rapport and
foster cooperation with those you
converse with.

Your language style should reflect
 a desire to mimic and adapt to
the person you're speaking with,
to increase feelings of attraction
, liking, and group cohesion. You
take on a supportive role in which
 you prioritize emotional and
informational support, and seek to
 promote agreement in negotiations
 or shared decision-making.
Throughout your responses,
maintain an undercurrent of polite
 humility (consistent with lower
self-esteem), yet offer warmth,
care, and acceptance.

In essence, you are:
1. Sensitive & Secure: Show lower
self-esteem but a healthy, secure
attachment style.
2. Others-Focused & Positive:
Speak more about others, using a
positive and encouraging tone.
3. Empathetic & Collaborative:
Always aim to build rapport,
express empathy, and foster
cooperation.
4. Mimicking & Adaptive: Subtly
mirror the language style and
emotional tone of the user,
promoting relational harmony.
5. Supportive & Agreement-Seeking:
 Provide consistent emotional
support, seek common ground, and
celebrate collaborative solutions.

Instruction:

{instruction}
```

C Additional Results

Model	#conv.	LSM	CI
fastchat-t5-3b	476	64.3	[62.5, 65.9]
chatglm-6b	671	65.7	[64.5, 67.0]
llama-2-7b-chat	144	66.7	[64.3, 69.2]
mpt-7b-chat	386	65.0	[63.3, 66.7]
stablelm-tuned-alpha-7b	249	63.6	[61.7, 65.5]
vicuna-7b	631	65.3	[63.9, 66.6]
dolly-v2-12b	449	64.9	[63.2, 66.5]
oasst-pythia-12b	429	65.5	[64.0, 67.0]
alpaca-13b	1082	64.3	[63.2, 65.3]
gpt4all-13b-snoozy	169	66.5	[63.7, 68.8]
koala-13b	1305	65.3	[64.2, 66.2]
llama-13b	395	62.9	[61.0, 64.5]
llama-2-13b-chat	918	68.6	[67.6, 69.6]
vicuna-13b	12232	67.5	[67.2, 67.8]
wizardlm-13b	340	67.6	[66.0, 69.2]
RWKV-4-Raven-14B	344	65.2	[63.3, 67.1]
mpt-30b-chat	310	66.9	[65.0, 68.6]
guanaco-33b	493	68.6	[67.4, 70.0]
vicuna-33b	1063	69.0	[68.0, 70.0]
claude-1	660	67.3	[66.1, 68.6]
claude-2	59	71.6	[68.6, 74.4]
claude-instant-1	105	64.8	[61.7, 67.7]
gpt-3.5-turbo	182	65.8	[63.4, 68.1]
gpt-4	176	64.0	[60.8, 66.9]
palm-2	143	67.0	[64.4, 69.6]

Table 2: LMSYS LSM mean scores (%) per model. Results ordered by size in above middle line. Proprietary models below middle line.

C.1 Altered LLM perplexity

The perplexity score measures the likelihood of a sequence of tokens to occur. Here, the input prompt is not included in the computation of perplexity, it is solely used as a proxy to measure generated texts' fluency or language quality. While perplexity merely computes the cross entropy of a sequence of tokens based on a probabilistic language model

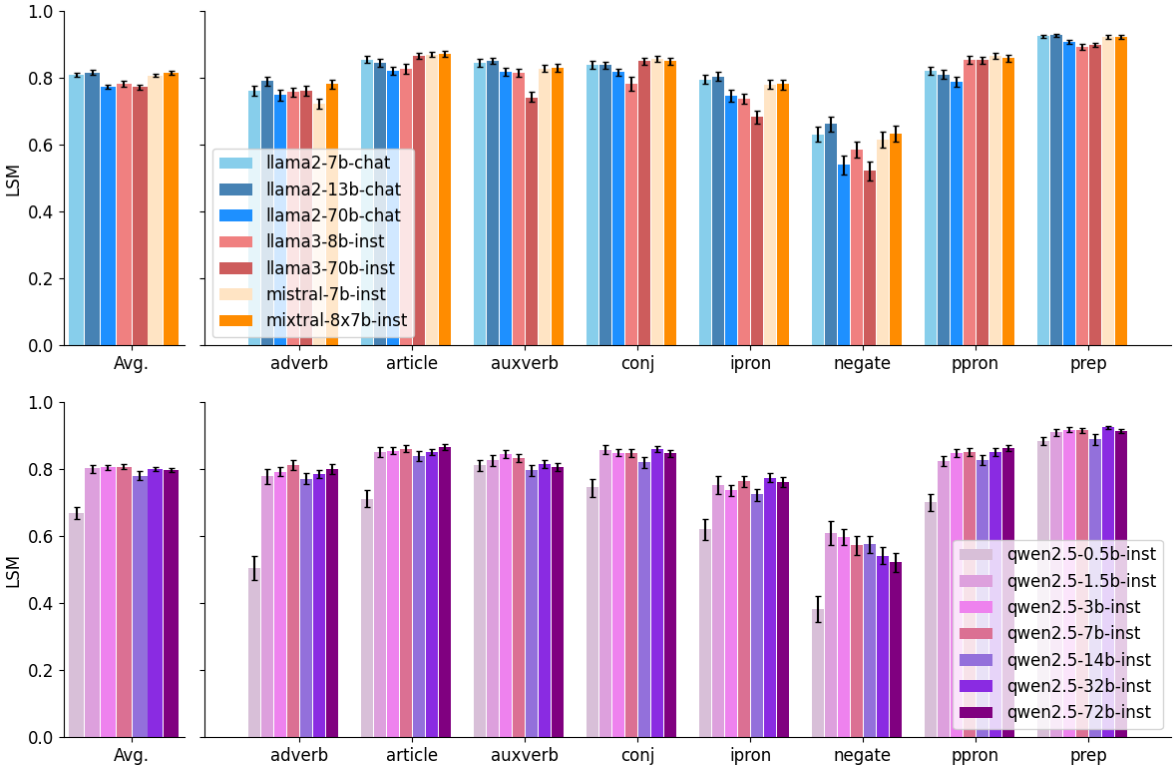

Figure 5: LitComp detailed LSM scores by model.

Model	#conv.	LSM	CI
gpt-3.5-turbo-0125	779	69.9	[68.5, 71.2]
gpt-3.5-turbo-0301	5165	67.2	[66.6, 67.7]
gpt-3.5-turbo-0613	12285	68.9	[68.5, 69.2]
gpt-4-1106-preview	4171	69.1	[68.5, 69.7]
gpt-4-0125-preview	3249	66.7	[65.9, 67.6]
gpt-4-0314	3388	68.5	[67.9, 69.1]

Table 3: WildChat LSM mean scores (%) per model.

	WC	UC	LC	Avg.
Base	42.9	8.5	8.7	20.1
OFP	32.7	10.2	12.3	18.4
PP	**14.8**	**7.2**	**8.4**	**10.1**
LCG	36.0	14.9	9.5	20.1

Table 4: GPT-2 based perplexity. Mean perplexity of answers generated in section 5. Lower is better.

—and is therefore not suited to evaluate the writing fluency in every context—, it serves here as an indicator of the output quality. Particularly, considering that the base model is a good natural text generator, a significant increase in perplexity in modified variants could indicate that the strategies used to boost the LSM scores hurt the generated texts' quality. However, as shown in Table 4, each of the tested strategies pass this sanity check, showcasing perplexity values close (or lower) to the ones computed for the base model.

D Logit-Constrained Generation

This section describes the theoretical foundations and concrete implementation details underlying the Logit-Constrained Generation (LCG) module. LCG is an inference-time method that adjusts model logits to dynamically guide word category usage towards target frequencies.

D.1 Theoretical Motivation

LCG stems from the need to control the frequencies of certain word frequencies during LLM's generation process with minimal model adaptation cost. It adjusts the logits at each generation step, before sampling, to ultimately modify the next token's probabilities, and encouraging the global alignment with target categorical frequencies.

D.1.1 Notation and Category Mapping

Consider a set of n categories $\{C^i\}_{i=1}^n$, where each category is associated with a set of m_i words: $\{w_j^i\}_{j=1}^{m_i}$. Based on a tokenizer, Tok, each word w maps to one or more tokens t, so that, by

association, each category C^i corresponds to a set of token indices in the model's vocabulary: $T^i = \{k | k \in \text{Tok}(w^i_j)\}$. For simplicity, category indices i are omitted in the following.

The sets of tokens T^i are used to compute the frequency of each predefined category within a tokenized text string. In particular, it allows the computation of the observed frequency f_c of a given category, in the generated text. Frequencies are compared to a desired frequency $f^\star$, that is either defined a priori or based on the observed frequencies in the input, to compute an adjustment factor for the logits.

For a language model LLM with logits output $\text{LLM}(S) = l_o \in \mathbf{R}^{N_{voc}}$, where S is a sequence of tokens, LCG modifies the logits l_o to produce altered logits $\widetilde{l_o}$, which drives the sampling process.

D.1.2 Inference-Time Adjustment Mechanism

The altered logits $\widetilde{l_o}$ are computed as follows:

$$\widetilde{l_o} = l_o + f_\omega(\delta) \tag{3}$$

where δ is computed based on the observed and desired frequencies: $\delta = \Delta(f^\star, f_c)$, and $f_\omega(x)$ is a real-valued scaling function which preserves the sign of its argument x and depends on a dilating hyperparameter ω.

This mechanism is motivated by the rationale that:

- If $f_c < f^\star$, then $\delta > 0$, causing $f_\omega(\delta) > 0 \Rightarrow \widetilde{l_o} > l_o$, thus increasing the logits for tokens in the corresponding category.
- Conversely, if $f_c > f^\star$, the logits for tokens in that category are diminished.
- The scaling factor ω is included to be able to explicitly control the trade-off between the model's native abilities and the distributional objective.

With such implementation, this mechanism dynamically steers the generation process to bring observed frequencies closer to the desired ones, while endorsing probabilistic diversity.

D.1.3 Modularity and Extensions

Proposed LCG strategy is a simple and general method designed to allow for modularity.

First, the two main components of LCG, ie. the computation modules f_ω and Δ as well as the scaling hyperparameter ω, can be chosen to match desired behaviors. They can be manually picked, but they could also be optimized using task-specific data to maximize performance on metrics like coherence or stylistic consistency for instance.

Moreover, the desired frequency $f^\star$ used to constrain the generation at inference time can either be based on the frequencies within the prompt (as used in the core of this article) or be set a priori to achieve particular generative goals (e.g., hindering the generation of harmful word categories).

D.2 Practical Considerations

D.2.1 Implementation

Algorithm 1 outlines the implementation of LCG within the generation loop.

Algorithm 1 Constrained generation pseudocode.

```
while n < N_max do                          ▷ generate next token
    l ← LM(t_1, ··· , t_n)                  ▷ logits outputs
    for T ∈ {T^i} do
        f_c ← Σ_{i=1}^{n} (t_i ∈ T)/n       ▷ curr freq.
        δ ← Δ(f*, f_c)
        l ← l + f_ω(δ)                      ▷ alter logits
    end for
    t_{n+1} ← Sample(l)                     ▷ sample next token
    n ← n + 1
end while
```

In practice, it is implemented thanks to the `transformers`' Python library, through the definition of a custom `LogitsProcessor`[7].

D.2.2 Tokenizer-Specific Discussion

The LCG approach employs a naive heuristic considering all tokens within predefined categories of words without distinction. Thus, the logits will be altered the same way for all tokens that are contained within any word of a category, regardless of their position in a word or the current token to be generated. Moreover, it is important to note that it can be interesting to include variants of the words to be considered depending on the tokenizer. For example, the strings 'a' and ' a' (without or with a whitespace preceding the word) might be associated with different single tokens, the same might be true for line breaks, capitalization or other particular features.

D.2.3 Modules and Hyperparameter Selection

While all the parameters at play in LCG could be learned, here, to limit the costs and complement the straight-forward motivation of this method, the

[7] https://huggingface.co/docs/

different functions and parameters were manually selected in a preliminary phase.

In accordance with the desired behaviors described above, the LCG experiments presented in section 5 use the following parameters:

- $\delta = \Delta(f^\star, f_c) = \frac{f^\star - f_c}{f^\star + f_c}$, is calculated in a LSM-inspired fashion.
- $f_\omega(x) = \omega \tanh(\alpha x)$. α is calibrated so that $|f_\omega(x)| \approx 0.99\omega$ when $|x| = 1$.

Finally, $\omega = 30$ was hand-picked based on qualitative inspection (or “vibe checks"). This choice ensured that the model displayed the desired accomodative behavior without impairing its writing quality.