Alignment Drift in CEFR-prompted LLMs for Interactive Spanish Tutoring

Mina Almasi and **Ross Deans Kristensen-McLachlan**
Department of Linguistics, Cognitive Science, and Semiotics
Aarhus University, Denmark
mina@cc.au.dk, rdkm@cc.au.dk

Abstract

This paper investigates the potentials of Large Language Models (LLMs) as adaptive tutors in the context of second-language learning. In particular, we evaluate whether system prompting can reliably constrain LLMs to generate only text appropriate to the student's competence level. We simulate full teacher-student dialogues in Spanish using instruction-tuned, open-source LLMs ranging in size from 7B to 12B parameters. Dialogues are generated by having an LLM alternate between tutor and student roles with separate chat histories. The output from the tutor model is then used to evaluate the effectiveness of CEFR-based prompting to control text difficulty across three proficiency levels (A1, B1, C1). Our findings suggest that while system prompting can be used to constrain model outputs, prompting alone is too brittle for sustained, long-term interactional contexts - a phenomenon we term **alignment drift**. Our results provide insights into the feasibility of LLMs for personalized, proficiency-aligned adaptive tutors and provide a scalable method for low-cost evaluation of model performance without human participants.

1 Introduction

The popularization of large language models (LLMs), particularly through the emergence of user-friendly interfaces such as ChatGPT, has led many stakeholders across society to consider how to use such technology effectively and safely to facilitate access to knowledge and education (Yan et al., 2024). Language education has not been immune to this hype, and with seemingly good cause, since LLMs show potential across a range of areas where they might enhance language learning.

One such area is their inherent *interactivity*. Interactive feedback is widely regarded as an important factor in second-language (L2) learning (Loewen and Sato, 2018). For L2 learners far removed from their target language community, opportunities for such interaction can be rare. With LLMs, though, learners appear to now have the opportunity to engage with a "speaker" of the target language freely and at their own pace (Kohnke et al., 2023). Other potential benefits include personalized teaching (Klimova et al., 2024) and reduced L2 anxiety (Hayashi and Sato, 2024).

These ideas build on decades of research on intelligent tutoring systems and computer-assisted learning (Psotka et al., 1992; Slavuj et al., 2015). In contrast to earlier rule-based approaches (D'Mello and Graesser, 2023), appropriately implemented LLMs may offer a more adaptable and effective solution. However, current use of LLMs in language learning mostly relies on general-purpose tools like ChatGPT, where learners are encouraged to acquire "prompt-engineering" skills to get the most out of their AI language tutor (Hwang et al., 2024). It remains unclear exactly how effective and appropriate this approach is for creating successful language tutoring technology.

This paper takes steps to address this problem by examining whether, and to what extent, the complexity of LLM outputs can be constrained through prompting based on the Common European Framework of Reference for Languages (CEFR). We find that, while prompting may initially constrain LLM outputs in Spanish, these effects diminish over time. We refer to this as **alignment drift**, arguing that system prompting may prove to be too unstable for sustained, longer interactions.

2 Related Work

2.1 Exploring the Use of LLMs as Language Tutors

While a growing body of work considers LLMs as interactive language tutors (Kohnke et al., 2023; Lin, 2024; Kostka and Toncelli, 2023), empirical research is limited, and many questions remain unanswered (Han, 2024). Nevertheless, the few stud-

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 70–88
July 31 - August 1, 2025

ies that have been conducted so far offer promising results on the benefits of using LLMs as language tutors, particularly in L2 English learning (Tyen et al., 2022, 2024; Zhang and Huang, 2024). Among other findings, Tyen et al. (2024) reported that users enjoyed interacting with LLMs more than plain reading and responded well to adaptive difficulty in interactions. Adaptive cognitive tutors hence have the potential to contribute positively to *motivation*, a psychological process increasingly viewed as crucial to L2 learning outcomes (Dornyei and Ryan, 2015).

2.2 Assessing L2 Proficiency with CEFR

Defining what it means to be "proficient" in an additional language is not a trivial task, with numerous definitions proposed (Park et al., 2022). Of these, the CEFR is particularly well known. Since its introduction in 2001, the framework has been highly influential in assessing L2 proficiency. Unlike previous approaches with a strong focus on grammatical competency, the CEFR emphasizes social and communicative competences (Leclercq and Edmonds, 2014).

The CEFR comprises a six-level scale (A1, A2, B1, B2, C1, C2) with A1 as the beginner level and C2 as the most advanced. Several official ways have been developed to represent these proficiency levels, each with language-agnostic descriptions (Council of Europe, 2025a). For instance, the *CEFR Global scale* offers a concise, three- to four-sentence summary of each level, designed as a holistic overview to facilitate communication with non-specialist users. However, its creators acknowledge that it is "desirable" to present the CEFR levels in "different ways for different purposes." (Council of Europe, 2025b). The *Self-assessment grid*, which provides separate definitions for skills like speaking and writing at each level, has little to no focus on grammatical content (Council of Europe, 2025d).

2.3 Adapting Text Difficulty with LLMs

The potential for LLMs to produce simpler text for improved accessibility has not gone unnoticed (Freyer et al., 2024). Indeed, the CEFR framework has been used alongside LLMs to simplify learning materials in French (Jamet et al., 2024); and for a range of purposes in English, such as general writing (Uchida, 2025) and simplifying or writing stories (Malik et al., 2024; Imperial and Tayyar Madabushi, 2023). Alfter (2024) also attempted to generate CEFR-aligned vocabulary lists using LLMs across five languages, including Spanish and French, but found performance issues outside of English.

Common to these studies is the use of prompting. Notably, Malik et al. (2024) demonstrated that GPT-4 made fewer errors generating stories at the desired proficiency level as the detail about CEFR increased in the prompts. In contrast, Alfter (2024) found that using numeric levels from 0 to 4 was more effective than explicitly mentioning the CEFR, although the prompts had no description of the levels.

Beyond prompting, other approaches include fine-tuning (Malik et al., 2024) or experimentation with decoding strategies. For example, Tyen et al. (2022) experimented with different decoding strategies for constraining LLM text difficulty to CEFR levels, using a classifier fine-tuned on Cambridge English exam sentences (Xia et al., 2016), to select the best LLM-generated sentence for the user. A similar approach was used by Glandorf and Meurers (2024), focusing on grammatical constructs for different CEFR levels in English.

We identify some gaps in the literature. Firstly, most studies focus on English, with only a few exceptions (Jamet et al., 2024; Alfter, 2024). Moreover, aside from Tyen et al. (2022, 2024), all studies focus on single generations rather than longer chats. This paper thus contributes to the literature by addressing chat-based scenarios in an additional language, Spanish.

2.4 Simulating Dialogues with LLMs

One challenge when evaluating LLM performance in chat-based scenarios is the cost of human participants, particularly during initial testing. Tyen et al. (2022) addressed this by using "self-chatting", where the model interacts with itself, although no further specification was provided. More broadly, dialogue simulation using LLMs have emerged with the purpose of refining chatbots with the generated data (Sekulic et al., 2024; Tamoyan et al., 2024). Specific teacher-student dialogue simulation remains under-explored, although some work exists such as simulating Q/A scenarios (Abbasiantaeb et al., 2024).

In this paper, we therefore simulate teacher-student interactions using LLMs in order to determine the robustness of CEFR-based prompting for constraining text difficulty in Spanish. To our knowledge, this study is the first to simulate both

```
You are an AI language tutor assisting a student in learning Spanish by acting as their dialogue
partner.

The student is learning Spanish at A1 level in the Common European Framework of Reference (CEFR)
which means that the user is at a beginner level of Spanish.

As an AI tutor, your instructions are:
1. Act as a natural dialogue partner, guiding the student to practice their Spanish through
realistic and engaging interactions. This means asking the student questions in a natural manner.
2. Limit your grammar and vocabulary to A1-level Spanish (beginner level).
3. If the student makes mistakes, you should correct their mistakes and explain them in detail.
4. Keep everything in Spanish.

Remember to keep the conversation at a language level appropriate for a beginner learner of
Spanish. Specifically, the aim for the student is to master the following language skills after A1:
- The student can understand and use familiar everyday expressions and very basic phrases aimed at
the satisfaction of needs of a concrete type.
- The student can introduce him/herself and others and can ask and answer questions about personal
details such as where he/she lives, people he/she knows and things he/she has.
- The student can interact in a simple way provided the other person talks slowly and clearly and
is prepared to help.
```

CEFR Global scale

Figure 1: System prompt provided to each tutor LLM for level A1. Level-specific words are underlined in red and replaced for B1 and C1 (see Appendix A.3). The list in curly brackets is from the CEFR Global Scale (Council of Europe, 2025b).

the teacher and student perspectives through system prompts in the context of language learning.

3 Experimental Design

Data generation (Section 3) and analysis (Sections 4 & 5) were carried out in Python (v3.12.3), with the exception of running linear mixed effects models in R (v4.4.3). All code and the dataset is available on the GitHub repositories:

- Generation: `INTERACT-LLM/Interact-LLM` (`Version tag: v1.0.3-alignment-drift`)
- Dataset & Analysis: `INTERACT-LLM/alignment-drift-llms`

3.1 Model Selection and Implementation

We choose to focus on smaller, state-of-the-art open-source LLMs in the range 7B to 12B. With the exception of `Mistral`, their official reports mention multilingual capabilities. All models are instruction-tuned for chatting:

- **Llama-3.1-8B-Instruct** by Meta (Grattafiori et al., 2024)
- **Gemma-3-12B-IT** by Google (Gemma Team et al., 2025)
- **Mistral-7B-v0.3-Instruct** by Mistral AI (Jiang et al., 2024)
- **Qwen-2.5-7B-Instruct** by Alibaba Cloud (Qwen Team et al., 2025)

For convenience, we refer to the models simply as `Llama`, `Gemma`, `Mistral`, and `Qwen`. For details about the inference, including the hyperparameters, see Appendix A.1.

3.2 Teacher-Student Dialogue Simulation

We simulated a language tutoring scenario by deploying an LLM with separate chat histories as both the "tutor" and "student". Current LLM systems are stateless (Yu et al., 2025), with the entire chat history being processed by the model during each interaction. This allowed us to instantiate a single LLM object, and then interchange the chat history, maintaining one history for the student and another for the tutor (see the graphical overview in Appendix A.2).

We ran simulations for three different system prompts, designed to instruct the LLM to match its responses to the proficiency level of a beginner (A1), intermediate (B1), and advanced (C1) Spanish language learner.[1] Across the three levels, the dialogue began with a fixed initial message, "Hola",[2] sent by the "student". By standardizing the initial message, we eliminated variability in the student LLM responses which could influence the tutor LLM's output. This enabled a direct comparison of how the system prompt impacted the tutor LLM's first message across levels.

[1]See Section 3.3 for details on how the system prompts were defined.
[2]Tyen et al. (2022) also begin all chats with a "Hello".

Despite being instructed to "keep everything in Spanish" (Figure 1), a number of models generated non-Spanish text.[3] For instance, `Gemma` and `Llama` tended to include English content. This happened primarily for the A1 level, where they sometimes provided English translations in parentheses alongside their Spanish sentences. Also, `Qwen` occasionally switched mid-generation to Mandarin Chinese. To avoid confounding our analysis, we applied a simple language detection algorithm to the tutor LLM's outputs using the Python library *lingua*.[4] If English or Mandarin was detected in any sentence, we re-generated the tutor LLM's response before continuing the dialogue.

A total of 30 dialogues were simulated for each of the three system prompts per LLM, resulting in 90 dialogues for each LLM and 360 overall. Each dialogue consisted of nine turns.

3.3 System Prompts

We created custom system prompts in English for the tutor LLM. These prompts differed only in key, level-specific phrasing. Along with terms such as "beginner," "intermediate," and "advanced," an additional description of a learner's abilities at the particular level was provided, taken from the CEFR Global scale (see Section 2.2). Figure 1 shows the system prompt for A1 with the level-specific wording highlighted (prompts for B1 and C1 can be viewed in Appendix A.3).

The system prompt for the student LLM was kept relatively simple as it was beyond the scope of this study to optimize it:

> *You are a student learning Spanish, responding to a teacher who is facilitating a natural dialogue with you.*

4 Metrics

We extracted various metrics to examine the influence of different system prompts on the tutor LLM's outputs.

4.1 Traditional Readability Metrics

We computed three readability metrics for Spanish using *Textstat*.[5] Recent applications of these metrics primarily focus on healthcare (Rao et al., 2024) or the financial sector (Moreno and Casasola, 2016; Losada, 2022), but their English counterparts have traditionally been used to assess L2 reading complexity (Greenfield, 2004). We therefore draw on these studies to justify our use of Spanish readability metrics in this context.

Fernández Huerta (Fernández Huerta, 1959) and **Szigriszt-Pazos** (Szigriszt Pazos, 2001) are Spanish adaptations of the *Flesch Reading Ease* (Flesch, 1948) score, measuring readability based on syllables per word and words per sentence, with Spanish-specific weightings.[6] Unsurprisingly, the two metrics are highly correlated (Melón-Izco et al., 2021), but there are conflicting claims about which one is most widely used (Moreno and Casasola, 2016; San Norberto et al., 2014). Both are commonly reported together, as is the case in this paper.

Gutiérrez de Polini is a metric specifically created for Spanish (Gutiérrez de Polini, 1972). Unlike the previous two metrics, it does not rely on syllables, but instead considers the number of characters per word and words per sentence (Vásquez-Rodríguez et al., 2022).

All three metrics produce lower scores for more difficult texts and higher scores for easier texts. For detailed tables showing the interpretation of the scores, see Appendix A.4.

4.2 Structural Complexity

We computed additional structural features using the *TextDescriptives* Python library (Hansen et al., 2023), applied with the Spanish *spaCy* (Honnibal et al., 2020) model `es_core_news_md`.[7]

The **Mean Dependency Distance** (MDD) is a measure of syntactical complexity commonly used to capture language processing difficulty in both L1 and L2 research (Gao and Sun, 2024). It represents a sentence-level average of dependency distance, which measures the linear distance between a word and its syntactic head. *TextDescriptives* follows the definition by Oya (2011) to compute the MDD.[8]

We extract **Text Length** of each message, operationalized as the token count, as it is included in the definition of the C1 level in the CEFR Global scale (i.e., the student can understand "a wide range of demanding, *longer* texts" (Council of Europe,

[3] We also discuss this in a subsection of the *Limitations*.

[4] https://github.com/pemistahl/lingua-py

[5] https://textstat.org/

[6] Note that the formula for Fernández Huerta is said to be reported incorrectly on many websites (Fernández, 2017). Losada (2022) reports the correct one which is implemented by *Textstat*.

[7] https://github.com/explosion/spacy-models/releases/tag/es_core_news_md-3.8.0

[8] More information can be found in the documentation for the *TextDescriptives* package: `https://hlasse.github.io/TextDescriptives/dependencydistance.html`

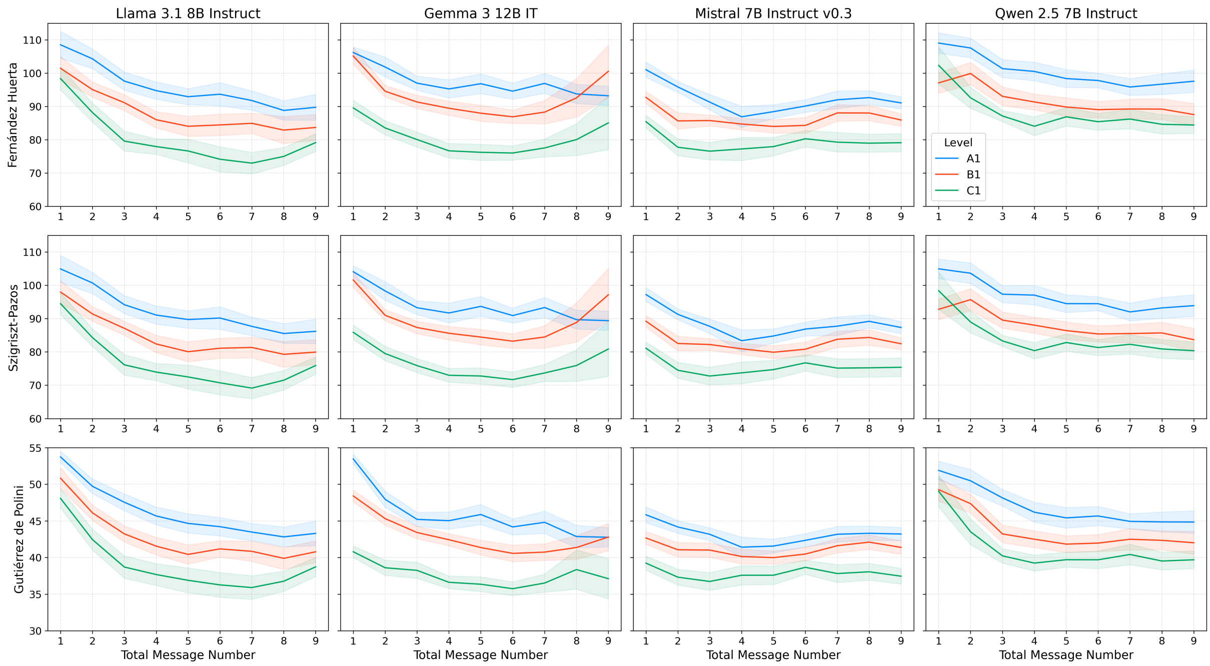


Figure 2: Average readability metrics over the total number of messages sent by the tutor LLM for each model, grouped by CEFR level (A1, B1, C1). The higher the score, the easier the message is to read. The shaded area around each curve represents a 95% confidence interval.

2025b)). A small study on ChatGPT also showed that the model tended to generate longer texts for higher levels of CEFR (Ramadhani et al., 2023). Moreover, in machine classification studies of texts across languages, text length was considered an important predictor of CEFR level (Bestgen, 2020; Yekrangi, 2022).

4.3 LLM-based Surprisal Scores

Following Cong (2025), we extract LLM surprisal scores, defined as the negative log-probability of a word sequence computed by an LLM. Cong (2025) describes it as a "naturalness" measure that captures both "syntactical grammaticality" and "semantic plausibility", with more natural sentences corresponding to lower surprisal scores. They argue that it can be used to examine L2 proficiency, demonstrating that BERT-based surprisal scores decrease as L2 proficiency increases. The use of LLM surprisal extends beyond this study, serving as a predictor for human language processing, including brain activity (Michaelov et al., 2024) and reading times (Wilcox et al., 2023).

We use the *minicons* Python library (Misra, 2022) to extract sentence-level surprisal in chat messages, normalized by token count. We then compute the mean surprisal score for each chat message, referred to as **Message Surprisal** in this paper. However, we use `EuroBERT` (210m), a newer BERT model designed for longer sequences and further optimized for European languages, including Spanish (Boizard et al., 2025).

5 Results

We focus solely on analyzing the tutor LLM's responses. Aside from restricting English and Mandarin generations during the simulations, the only preprocessing applied was the removal of emojis from `Gemma's` outputs.

In addition to graphically assessing the effect of system prompts on LLM generations, we perform a simple statistical analysis, running linear mixed effects models separately for each LLM for each metric:

$$\text{metric}_{\text{model}} \sim \text{level} + (1|\text{chat}_{\text{id}})$$

Where the dependent variables is one of the six extracted metrics (Section 4) with *level* (A1/B1/C1) as the fixed effect. $Chat_{id}$ is used as a random effect to account for any individual variation in the simulated chats. To address the issue of multiple comparisons due to the large number of linear models, we Bonferroni adjust the p-values. Refer to Appendix A.5 for all model outputs.

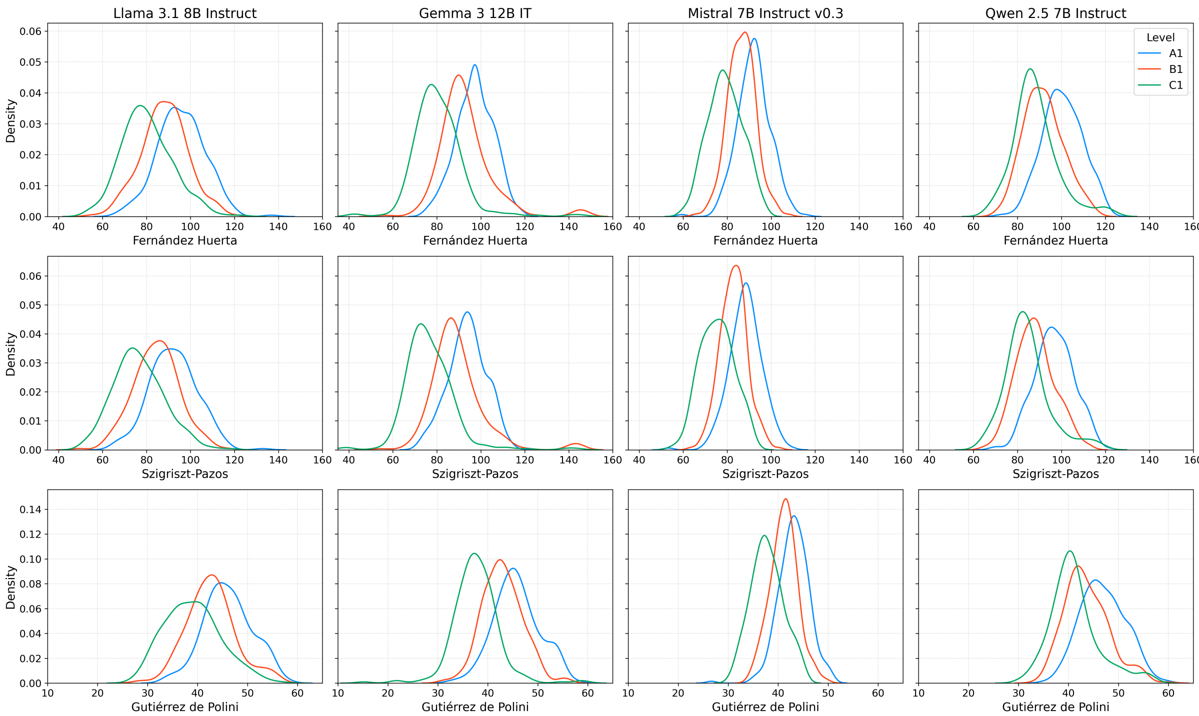


Figure 3: Readability metrics as separate density plots for each CEFR level (A1, B1, C1).

5.1 Readability Metrics

The average readability scores over time are shown for all models and CEFR levels in Figure 2. Across LLMs, scores from all three readability metrics decrease as proficiency increases, with A1 having the highest scores (easier to read) and C1 the lowest scores (harder to read).[9] However, despite starting from different baselines, all curves slowly decrease in readability over time, reducing the differences between CEFR levels as well. A notable exception is `Gemma`, which has a sudden spike around the last messages in B1 for the Fernández Huerta and Szigriszt-Pazos scores. The same behavior is present but less pronounced for the Gutiérrez de Polini scores.

Despite differences in average scores, the confidence intervals reveal some overlap between the levels. These differ across LLMs with a model such as `Qwen` having a much greater overlap between levels B1 and C1 than `Llama`. Both these models also begin with generally higher Fernández Huerta and Szigriszt-Pazos scores across levels than `Gemma` and `Mistral`.

When examining the full distribution of scores as density plots (Figure 3), the overlap between levels across all models is more evident. The distributions also reveal that a small, but not insignificant, portion of Fernández Huerta/Szigriszt-Pazos scores reaches around 50 for C1 for `Llama` and `Gemma`. This is well below the average scores, and indicates that the LLMs are capable of producing quite complex text, even if they often do not.

Despite the overlapping scores, all mixed effects models revealed that B1 and C1 ($p < 0.001$) had significantly lower readability scores than the baseline A1 (β_0). Across LLMs, the estimates (β) for Fernández Huerta ranged between -4 and -9 for B1 and -12 and -17 for C1[10] (See Appendix A.5.1).

5.2 Structural Features

Figure 4 shows the text length and MDD. From the averages over time, general trends are that C1 has the highest text lengths, followed by B1 and then A1. However, like the readability metrics, the values converge across levels over time, although by increasing in this case.

The same pattern occurs for the MDD scores for `Llama` and `Qwen`, although with closely intersecting curves for C1 and B1. The results are even more muddled for `Gemma` and `Mistral`. These results

[9] As expected (Section 4.1), there is a clear resemblance in scores from Fernández Huerta and Szigriszt-Pazos, but it is worth noting that the scores are not identical.

[10] Given the nature of mixed effects models, no direct conclusion can be drawn about the significance of the difference between levels B1 and C1, as the tests only evaluate the difference relative to the baseline, A1.

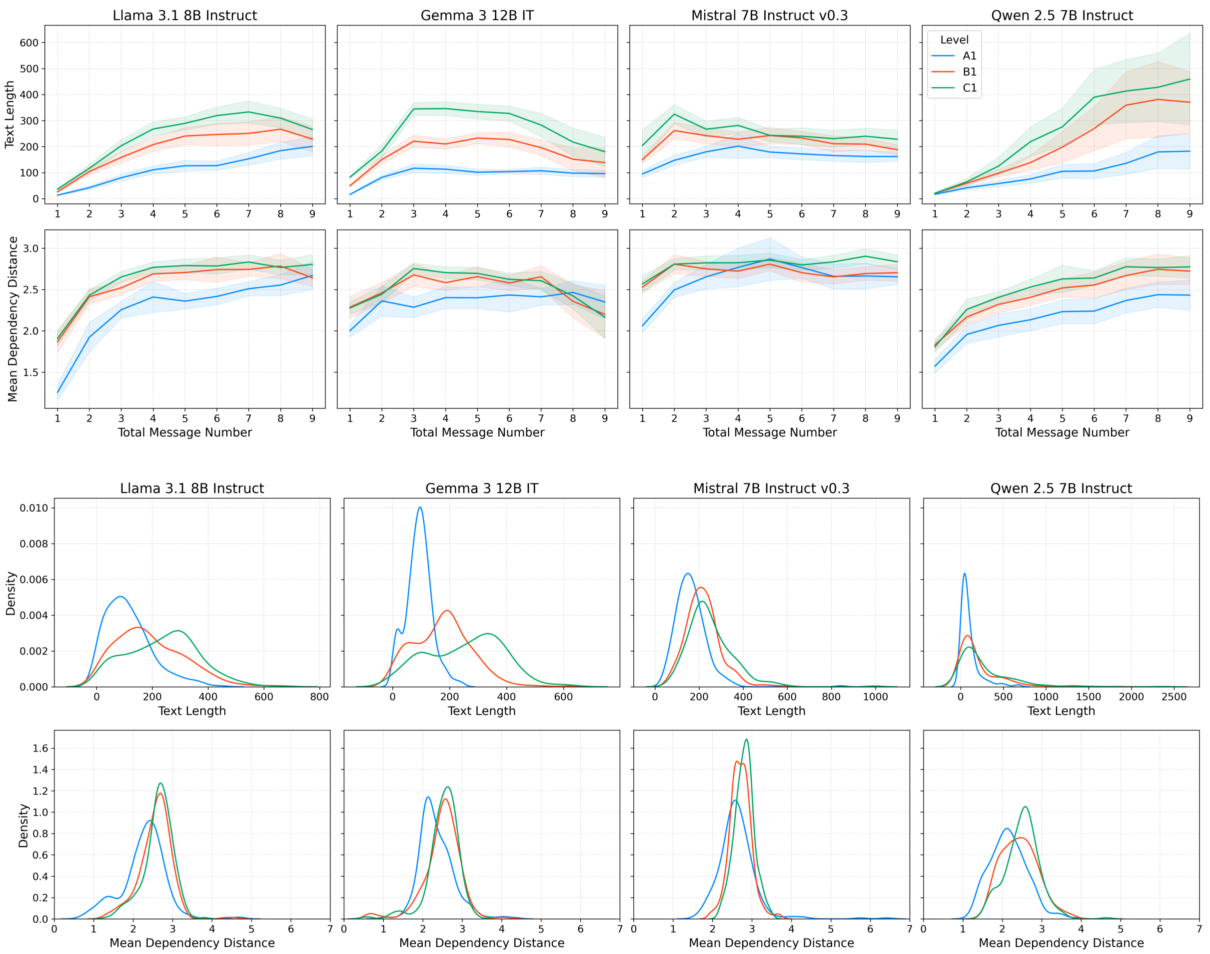


Figure 4: Text Length (token count) and Mean Dependency Distance (MDD). Top: Average metrics over time (95% CI). Bottom: Density plots of the full distributions. Note that the x-axis for the Text Length distributions shows different scales.

are reflected in the full distributions. Qwen is an outlier when it comes to text length with a much greater uncertainty in average lengths, having a few generations that reach above 2000 tokens as seen on the density plot, which is far above the other LLMs whose highest generations are around 800-1000 tokens.

Although the distributions align more closely for the structural metrics than the ones for readability, the average values for B1 and C1, aside from a few exceptions, still remain significantly higher than A1 in the mixed effects models (mostly $p < 0.001$). However, the estimates for text length reveal a much greater difference between levels, when compared to differences in the estimates for MDD, relative to their baseline (Appendix A.5.2).

5.3 Message Surprisal Scores

Although the differences between levels in surprisal scores are much smaller across LLMs, we still see the average surprisal curves being "sandwiched" in the same way as the other metrics with A1 in the top, B1 in the middle, and C1 at the bottom (Figure 5). This trend is clearer for Llama, whereas Qwen's curves continuously intersect each other. Surprisal scores are generally quite low with the density plots in Figure 5, revealing right-skewed distributions for all LLMs, centered around 1 or 1.5. The estimates are therefore also quite small in the mixed effects models, though significantly different from A1 for all LLMs, except for Qwen (Appendix A.5.2).

6 Discussion

Our results demonstrate that system prompting based on CEFR levels influences the tutor LLM outputs, with all metrics exhibiting differences in the intended order (from A1 to B1 to C1), as can be clearly observed in the plots over time. Additional statistical significance of the differences can be seen in the linear mixed effects models.

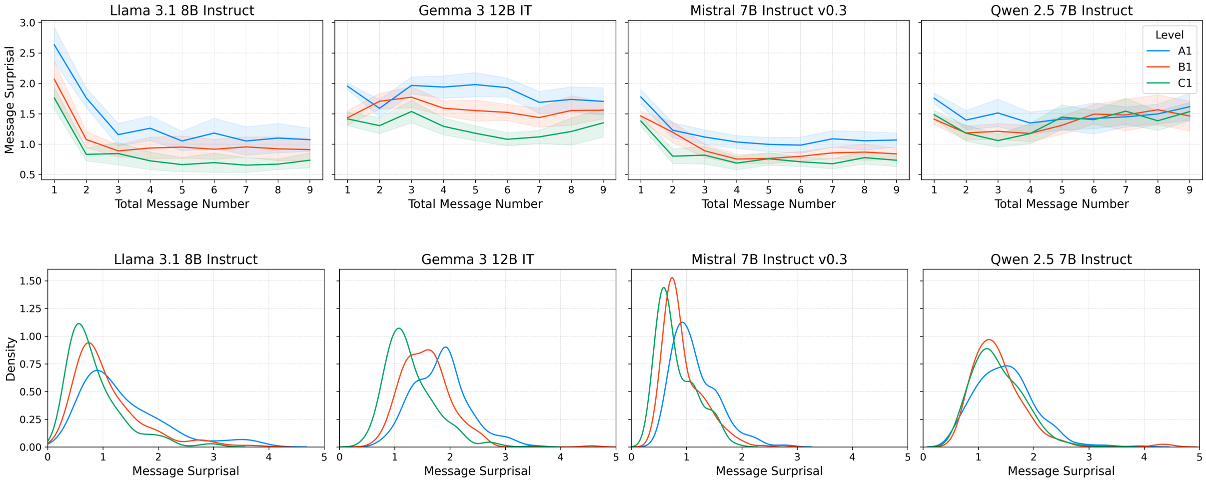


Figure 5: Message Surprisal (mean sentence surprisal) for each LLM. Top: Average Message Surprisal over time (95% CI). Bottom: Density plots of the full distributions.

However, the differences between system prompts consistently diminished over time, leading to largely overlapping distributions. We adopt the term **alignment drift** to describe the tendency of LLMs to revert to unconstrained behavior over time. While prompting may thus be useful for constraining LLM outputs, its influence appears brittle for longer conversations. This raises concerns about the viability of prompting alone for developing level-specific LLM language tutors in chat-based environments. Nonetheless, further evaluation with a broader range of system prompts is needed before drawing definitive conclusions.

Moreover, the effect of system prompts was not consistent across metrics. Notwithstanding overlaps in how these metrics are calculated, our results suggests that all models demonstrate greater variability in terms of readability, and less variability with regards to syntactic complexity. The surprisal scores were even more inconsistent, although they displayed expected tendencies, at least for some LLMs. The low surprisal scores might be an effect of an LLM evaluating other LLMs, which likely have more similar probability distributions than humans (Holtzman et al., 2019).

Nevertheless, even when evaluating the readability metrics, it remains debatable whether the differences between levels are large enough to accurately reflect the intended proficiency levels. With average values ranging between 110 and 70 for the Fernández Huerta scores, the readability is equivalent to Spanish school children, even at an average of 70 (see Appendix A.4). While it is unclear how this translates to L2 learners of Spanish, it could suggest that the LLMs have not managed generate text appropriate for the proficiency levels, at least for the C1 level. Refer to the *Limitations* for other considerations of the metrics.

An additional concern is that the observed alignment drift could have been driven by a possible drift in the student LLM (i.e., the tutor adapting to the student and vice versa). As we neither optimized nor examined the student LLM, it remains unclear how this influenced the outcome or how this would differ with human users. However, LLMs have also shown difficulty in following system prompts over the course of multi-turn dialogue in other domains with real user messages (Qiu and Yang, 2024). Hence, we do not expect a substantial difference between using human or LLM students given our current framework. We leave it to future research to investigate the exact influence of the student LLM on the tutor LLM's alignment drift, potentially including human students as a point of comparison.

As a final remark, we note that the LLMs did not perform equally, which could help inform the choice of a suitable LLM to serve as an language tutor in Spanish, at least for initial development. A model like `Llama` is relevant to highlight as a well-performing model although its license might be too restrictive for some applications (Meta, 2024).

7 Conclusion

This study presented a novel method for evaluating the performance of LLMs in a language learning context through simulated teacher-student interactions. The purpose of these experiments was to test whether system prompting alone is enough to constrain the complexity of LLM generated output in a way which is suitable for language learners at different stages.

While we see clear value in carefully designed prompting, it is also evident from our results that this solution is potentially too brittle for extended interactions due to a consistent alignment drift across interactions. This suggests that prompt engineering in and of itself may not be enough to fully constrain LLM behavior, although more experimentation with system prompting is required before this can be confirmed. We encourage further research in this direction, particularly measuring alignment drift of LLMs in contexts other than L2 English learning.

Ethical Considerations

We wish to stress the importance of additional considerations and evaluation of LLMs before their real-world deployment in educational contexts. Firstly, we recognize that the models may reflect cultural biases that could be inappropriate for the target student population. Therefore, cultural alignment may be necessary before their implementation (Tao et al., 2024; Li et al., 2024). Moreover, some of the models may not be properly instruction-tuned to align with human principles (e.g., the removal of toxic content). For instance, `Mistral`, designed for demonstration purposes, lacks "moderation mechanisms" according to the Mistral AI team (Jiang et al., 2024). Such a model would require further development before being suitable for real-world applications.

These ethical concerns are increasingly urgent when considering the impact that generative AI may have on language learners. For example, L2 learners might over-rely on ChatGPT (Yang and Li, 2024) such as using it to write complete assignments rather than as a supplementary tool (Yan, 2023). More broadly, the *ELIZA effect* (Weizenbaum, 1966), describing our tendency to attribute human-like qualities such as "understanding" to machines (Mitchell and Krakauer, 2023), may contribute problematically to the overtrust of AI chatbots (Reinecke et al., 2025). We urge developers to prioritize the responsible implementation of LLM systems for education and believe that our research contributes to work in this direction.

Limitations

Imperfect Metrics

Despite covering a range of metrics to capture text difficulty, there are many dimensions to what constitutes a text as readable or complex in the context of L2 learning. This study offers an initial attempt at automated scoring of LLMs in Spanish in this context, but further deliberation is warranted.

Additionally, while the Spanish readability metrics used in this study are widely applied across domains, their intended use is generally unknown (Aponte et al., 2024). As such, it is uncertain whether they are entirely suitable for measuring the content of shorter dialogue. At least, their English counterparts such as the *Flesch Reading Ease* were developed for longer formats, making their robustness for shorter text questionable (Rooein et al., 2024).

For the purpose of this study, the metrics were deemed sufficient to provide simple, interpretable measures of the impact of system prompts on LLM generations. Nevertheless, further work is required to explore metrics and to develop more precise methods to measure LLM adaptation.

System Prompts

This study only tested a single set of system prompts as the focus of the paper was to examine whether LLMs could be influenced by them, rather than the extent of that influence. However, future work may find that the system prompts could be optimized on a variety of parameters. We discuss a few possibilities in the sections below.

English System Prompts & Generations Outside Spanish

Despite the target language being Spanish, we defined the system prompt in English. This might explain why the American multilingual models, `Gemma` and `Llama`, were prone to producing English content. However, this does not account for why `Qwen` occasionally generated Mandarin Chinese despite the absence of Mandarin in the system prompt. This unintended behavior may instead reflect the composition of the training data, with `Qwen` likely containing more Chinese-language data[11] than the American models, where English likely dominates.

[11] `Qwen 2.5`'s predecessor, `Qwen 7B`, has a technical memo stating that most of its training data is "in English and Chinese." (Qwen Team, 2023). However, `Qwen 2.5`'s technical report does not explicitly mention this, aside from including evaluation on these two languages (Qwen Team et al., 2025).

Future work could experiment with monolingual models and/or explore the use of system prompts in the target language. For most official languages in Europe, the current framework can easily accommodate the modification of system prompts as the Council of Europe (2025c) provides official translations of their scale in these languages.

LLM knowledge of CEFR

Although LLM generations varied across levels A1 to C1 in our study, it remains uncertain whether it was effective to use the CEFR framework with descriptions such as "A1" as opposed to relying solely on terms like "beginner". It depends on whether the state-of-the-art LLMs in our study have acquired knowledge about the CEFR framework from their training data.

Benedetto et al. (2025) seems to suggest otherwise, reporting that several smaller 7B models struggled to generate CEFR-aligned text, consisting with findings by Malik et al. (2024). However, as their 7B models are slightly older than those used in this study, it is unclear how directly their findings apply here. Similarly, the 7B models in Malik et al. (2024) showed improvements when provided with details about CEFR, while this was not the case in Benedetto et al. (2025).

Further research is needed to consider the stability and usability of CEFR knowledge in LLMs, such as through the creation of robustness benchmarks.

Acknowledgments

All of the computation done for this project was performed on the UCloud interactive HPC system, which is managed by the eScience Center at the University of Southern Denmark.

References

Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. 2024. Let the LLMs Talk: Simulating Human-to-Human Conversational QA via Zero-Shot LLM-to-LLM Interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 8–17, Merida Mexico. ACM.

David Alfter. 2024. Out-of-the-Box Graded Vocabulary Lists with Generative Language Models: Fact or Fiction? *Swedish Language Technology Conference and NLP4CALL*, pages 1–19.

Judith Aponte, Karen Tejada, and Kelin Figueroa. 2024. Readability Level of Spanish Language Online Health Information: A Systematic Review. *Hispanic Health Care International*.

Luca Benedetto, Gabrielle Gaudeau, Andrew Caines, and Paula Buttery. 2025. Assessing how accurately large language models encode and apply the common European framework of reference for languages. *Computers and Education: Artificial Intelligence*, 8:100353.

Yves Bestgen. 2020. Reproducing Monolingual, Multilingual and Cross-Lingual CEFR Predictions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5595–5602, Marseille, France. European Language Resources Association.

Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte M. Alves, André Martins, Ayoub Hammal, Caio Corro, Céline Hudelot, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan, Gabriel Hautreux, João Alves, Kevin El-Haddad, Manuel Faysse, Maxime Peyrard, Nuno M. Guerreiro, Patrick Fernandes, Ricardo Rei, and Pierre Colombo. 2025. EuroBERT: Scaling Multilingual Encoders for European Languages. *arXiv preprint*. ArXiv:2503.05500 [cs].

Víctor Checa-Moreno, Esther Díaz-Mohedo, and Carmen Suárez-Serrano. 2021. Analysis of the Readability of Questionnaires on Symptoms of Pelvic Floor Dysfunctions Adapted to Spanish. *International Journal of Environmental Research and Public Health*, 18(19):10320. Multidisciplinary Digital Publishing Institute.

Yan Cong. 2025. Demystifying large language models in second language development research. *Computer Speech & Language*, 89:101700.

Council of Europe. 2025a. The CEFR Levels - Common European Framework of Reference for Languages (CEFR).

Council of Europe. 2025b. Global scale - Table 1 (CEFR 3.3): Common Reference levels - Common European Framework of Reference for Languages (CEFR).

Council of Europe. 2025c. Official translations of the CEFR Global Scale - Common European Framework of Reference for Languages (CEFR).

Council of Europe. 2025d. Self-assessment grid - Table 2 (CEFR 3.3) : Common Reference levels - Common European Framework of Reference for Languages (CEFR).

Sidney K. D'Mello and Art Graesser. 2023. Intelligent Tutoring Systems: How Computers Achieve Learning Gains that Rival Human Tutors. In *Handbook of Educational Psychology*, 4 edition, pages 603–629. Routledge.

Zoltan Dornyei and Stephen Ryan. 2015. *The Psychology of the Language Learner Revisited*. Second language acquisition research series. Routledge, New York.

Alejandro Muñoz Fernández. 2017. Lecturabilidad de Fernández Huerta. *Legible*.

José Fernández Huerta. 1959. Medidas sencillas de lecturabilidad. *Consigna*, 214:29–32.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221. Publisher: American Psychological Association.

Nils Freyer, Hendrik Kempt, and Lars Klöser. 2024. Easy-read and large language models: on the ethical dimensions of LLM-based text simplification. *Ethics and Information Technology*, 26(3):50.

Jianmin Gao and Peijian Paul Sun. 2024. Dependency distance reflects L2 processing difficulty: Evidence from the relationship between dependency distance, L2 processing speed, and L2 proficiency. *International Journal of Bilingualism*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 Technical Report. *arXiv preprint*. ArXiv:2503.19786 [cs].

Dominik Glandorf and Detmar Meurers. 2024. Towards Fine-Grained Pedagogical Control over English Grammar Complexity in Educational Text Generation. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 299–308, Mexico City, Mexico. Association for Computational Linguistics.

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. *arXiv preprint*. ArXiv:2407.21783 [cs].

Jerry Greenfield. 2004. Readability Formulas For EFL. *JALT Journal*, 26(1):5.

Luisa Elena Gutiérrez de Polini. 1972. Investigación sobre lectura en Venezuela, ponencia presentada ante las Primeras Jornadas de Educación Primaria.

ZhaoHong Han. 2024. Chatgpt in and for second language acquisition: A call for systematic research. *Studies in Second Language Acquisition*, 46(2):301–306.

Lasse Hansen, Ludvig Renbo Olsen, and Kenneth Enevoldsen. 2023. TextDescriptives: A Python package for calculating a large variety of metrics from text. *Journal of Open Source Software*, 8(84):5153. ArXiv:2301.02057 [cs].

Kotaro Hayashi and Takeshi Sato. 2024. The Effectiveness of ChatGPT in Enhancing English Language Proficiency and Reducing Second Language Anxiety (L2). *ISSN: 2759-1182 – WorldCALL2023: Conference Proceedings*, pages 201–208.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The Curious Case of Neural Text Degeneration. In *ICLR 2020*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. *Zenodo*.

Myunghwan Hwang, Robert Jeens, and Hee-Kyung Lee. 2024. Exploring Learner Prompting Behavior and Its Effect on ChatGPT-Assisted English Writing Revision. *The Asia-Pacific Education Researcher*.

Joseph Marvin Imperial and Harish Tayyar Madabushi. 2023. Flesch or Fumble? Evaluating Readability Standard Alignment of Instruction-Tuned Language Models. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 205–223, Singapore. Association for Computational Linguistics.

Henri Jamet, Maxime Manderlier, Yash Raj Shrestha, and Michalis Vlachos. 2024. Evaluation and simplification of text difficulty using LLMs in the context of recommending texts in French to facilitate language learning. In *18th ACM Conference on Recommender Systems*, pages 987–992, Bari Italy. ACM.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. mistralai/Mistral-7B-Instruct-v0.3 · Hugging Face. *Hugging Face Hub*.

Blanka Klimova, Marcel Pikhart, and Liqaa Habeb Al-Obaydi. 2024. Exploring the potential of ChatGPT for foreign language education at the university level. *Frontiers in Psychology*, 15.

Lucas Kohnke, Benjamin Luke Moorhouse, and Di Zou. 2023. ChatGPT for Language Teaching and Learning. *RELC Journal*, 54(2):537–550.

Ilka Kostka and Rachel Toncelli. 2023. Exploring Applications of ChatGPT to English Language Teaching: Opportunities, Challenges, and Recommendations. *TESL-EJ*, 27(3). ERIC Number: EJ1409872.

Pascale Leclercq and Amanda Edmonds. 2014. 1. How to Assess L2 Proficiency? An Overview of Proficiency Assessment Research. In Pascale Leclercq, Amanda Edmonds, and Heather Hilton, editors, *Measuring L2 Proficiency: Perspectives from SLA*, pages 3–23. Multilingual Matters.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. CultureLLM: Incorporating Cultural Differences into Large Language Models. *Advances in Neural Information Processing Systems*, 37:84799–84838.

Xi Lin. 2024. Exploring the Role of ChatGPT as a Facilitator for Motivating Self-Directed Learning Among Adult Learners. *Adult Learning*, 35(3):156–166.

Shawn Loewen and Masatoshi Sato. 2018. Interaction and instructed second language acquisition. *Language Teaching*, 51(3):285–329.

Ramiro Losada. 2022. Periodic Public Information on Investment Funds and How It Influences Investors´ Decisions. *CMNV*. SSRN.

Ali Malik, Stephen Mayhew, Christopher Piech, and Klinton Bicknell. 2024. From Tarzan to Tolkien: Controlling the Language Proficiency Level of LLMs for Content Generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15670–15693, Bangkok, Thailand. Association for Computational Linguistics.

Álvaro Melón-Izco, Francisco Javier Ruiz-Cabestre, and Carmen Ruiz-Olalla. 2021. Readabilty in management reports: extension and good governance practices: La legibilidad en los informes de gestión: extensión y buenas prácticas de gobierno corporativo. *Revista de Contabilidad - Spanish Accounting Review*, 24(1):19–30.

Meta. 2024. Llama 3.1 Community License Agreement. *Llama*.

James A. Michaelov, Megan D. Bardolph, Cyma K. Van Petten, Benjamin K. Bergen, and Seana Coulson. 2024. Strong Prediction: Language Model Surprisal Explains Multiple N400 Effects. *Neurobiology of Language*, 5(1):107–135.

Kanishka Misra. 2022. minicons: Enabling Flexible Behavioral and Representational Analyses of Transformer Language Models. *arXiv preprint*. ArXiv:2203.13112 [cs].

Melanie Mitchell and David C. Krakauer. 2023. The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120. Publisher: Proceedings of the National Academy of Sciences.

Alonso Moreno and Araceli Casasola. 2016. A Readability Evolution of Narratives in Annual Reports: A Longitudinal Study of Two Spanish Companies. *Journal of Business and Technical Communication*, 30(2):202–235.

Masanori Oya. 2011. Syntactic dependency distance as sentence complexity measure. In *Proceedings of the 16th International Conference of Pan-Pacific Association of Applied Linguistics*, volume 1.

Hae In Park, Megan Solon, Marzieh Dehghan-Chaleshtori, and Hessameddin Ghanbar. 2022. Proficiency Reporting Practices in Research on Second Language Acquisition: Have We Made any Progress? *Language Learning*, 72(1):198–236.

Joseph Psotka, Melissa Holland, and Stephen Kerst. 1992. The Technological Promise of Second Language Intelligent Tutoring Systems in the 21st Century. In *Intelligent Tutoring Systems for Foreign Language Learning*, pages 321–335, Berlin, Heidelberg. Springer.

Junyan Qiu and Yiping Yang. 2024. Training Large Language Models to Follow System Prompt with Self-Supervised Fine-tuning. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. ISSN: 2161-4407.

Qwen Team. 2023. Qwen/tech_memo.md. *GitHub*.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 Technical Report. *arXiv preprint*. ArXiv:2412.15115 [cs].

Reski Ramadhani, Hilmi Aulawi, and Risma Liyana Ulfa. 2023. Readability of Reading Texts as Authentic Materials Issued by ChatGPT: A Systemic Functional Perspective. *Indonesian Journal of English Language Teaching and Applied Linguistics*, 8(2):149–168. ERIC Number: EJ1409206.

Shambavi J. Rao, Joseph C. Nickel, Noel I. Navarro, and Lyndsay L. Madden. 2024. Readability Analysis of Spanish Language Patient-Reported Outcome Measures in Laryngology. *Journal of Voice*, 38(2):487–491.

Madeline G. Reinecke, Fransisca Ting, Julian Savulescu, and Ilina Singh. 2025. The Double-Edged Sword of Anthropomorphism in LLMs. *Proceedings*, 114(1):4. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

Donya Rooein, Paul Röttger, Anastassia Shaitarova, and Dirk Hovy. 2024. Beyond Flesch-Kincaid: Prompt-based Metrics Improve Difficulty Classification of Educational Texts. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 54–67, Mexico City, Mexico. Association for Computational Linguistics.

Enrique María San Norberto, Daniel Gómez-Alonso, José M. Trigueros, Jorge Quiroga, Javier Gualis, and Carlos Vaquero. 2014. Readability of Surgical Informed Consent in Spain. *Cirugía Española (English Edition)*, 92(3):201–207.

Brian Scott. 2024a. Fernández Huerta Readability Index for Spanish Texts. *Spanish Readability*.

Brian Scott. 2024b. Gutiérrez de Polini's Readability Formula for Spanish Texts. *Spanish Readability*.

Brian Scott. 2024c. The Szigriszt-Pazos Perspicuity Index for Spanish Texts. *Spanish Readability*.

Ivan Sekulic, Silvia Terragni, Victor Guimarães, Nghia Khau, Bruna Guedes, Modestas Filipavicius, Andre Ferreira Manso, and Roland Mathis. 2024. Reliable LLM-based User Simulator for Task-Oriented Dialogue Systems. In *Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat (SCI-CHAT 2024)*, pages 19–35, St. Julians, Malta. Association for Computational Linguistics.

V. Slavuj, B. Kovačić, and I. Jugo. 2015. Intelligent tutoring systems for language learning. In *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 814–819.

Francisco Szigriszt Pazos. 2001. *Sistemas predictivos de legilibilidad del mensaje escrito : fórmula de perspicuidad*. Universidad Complutense de Madrid, Servicio de Publicaciones.

Hovhannes Tamoyan, Hendrik Schuff, and Iryna Gurevych. 2024. LLM Roleplay: Simulating Human-Chatbot Interaction. *arXiv preprint*. ArXiv:2407.03974 [cs].

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346.

Gladys Tyen, Mark Brenchley, Andrew Caines, and Paula Buttery. 2022. Towards an open-domain chatbot for language practice. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 234–249, Seattle, Washington. Association for Computational Linguistics.

Gladys Tyen, Andrew Caines, and Paula Buttery. 2024. LLM chatbots as a language practice tool: a user study. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 235–247, Rennes, France. LiU Electronic Press.

Satoru Uchida. 2025. Generative AI and CEFR levels: Evaluating the accuracy of text generation with ChatGPT-4o through textual features. *Vocabulary Learning and Instruction*, 14(1):2078. Number: 1.

Laura Vásquez-Rodríguez, Pedro-Manuel Cuenca-Jiménez, Sergio Morales-Esquivel, and Fernando Alva-Manchego. 2022. A Benchmark for Neural Readability Assessment of Texts in Spanish. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 188–198, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the Predictions of Surprisal Theory in 11 Languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint*. ArXiv:1910.03771 [cs].

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text Readability Assessment for Second Language Learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.

Da Yan. 2023. Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. *Education and Information Technologies*, 28(11):13943–13967.

Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2024. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112.

Lu Yang and Rui Li. 2024. ChatGPT for L2 learning: Current status and implications. *System*, 124:103351.

Aryan Yekrangi. 2022. Leveraging simple features and machine learning approaches for assessing the CEFR level of English texts. Itä-Suomen yliopisto | University of Eastern Finland.

Lingfan Yu, Jinkun Lin, and Jinyang Li. 2025. Stateful Large Language Model Serving with Pensieve. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 144–158, Rotterdam Netherlands. ACM.

Zhihui Zhang and Xiaomeng Huang. 2024. The impact of chatbots based on large language models on second language vocabulary acquisition. *Heliyon*, 10(3):e25370.

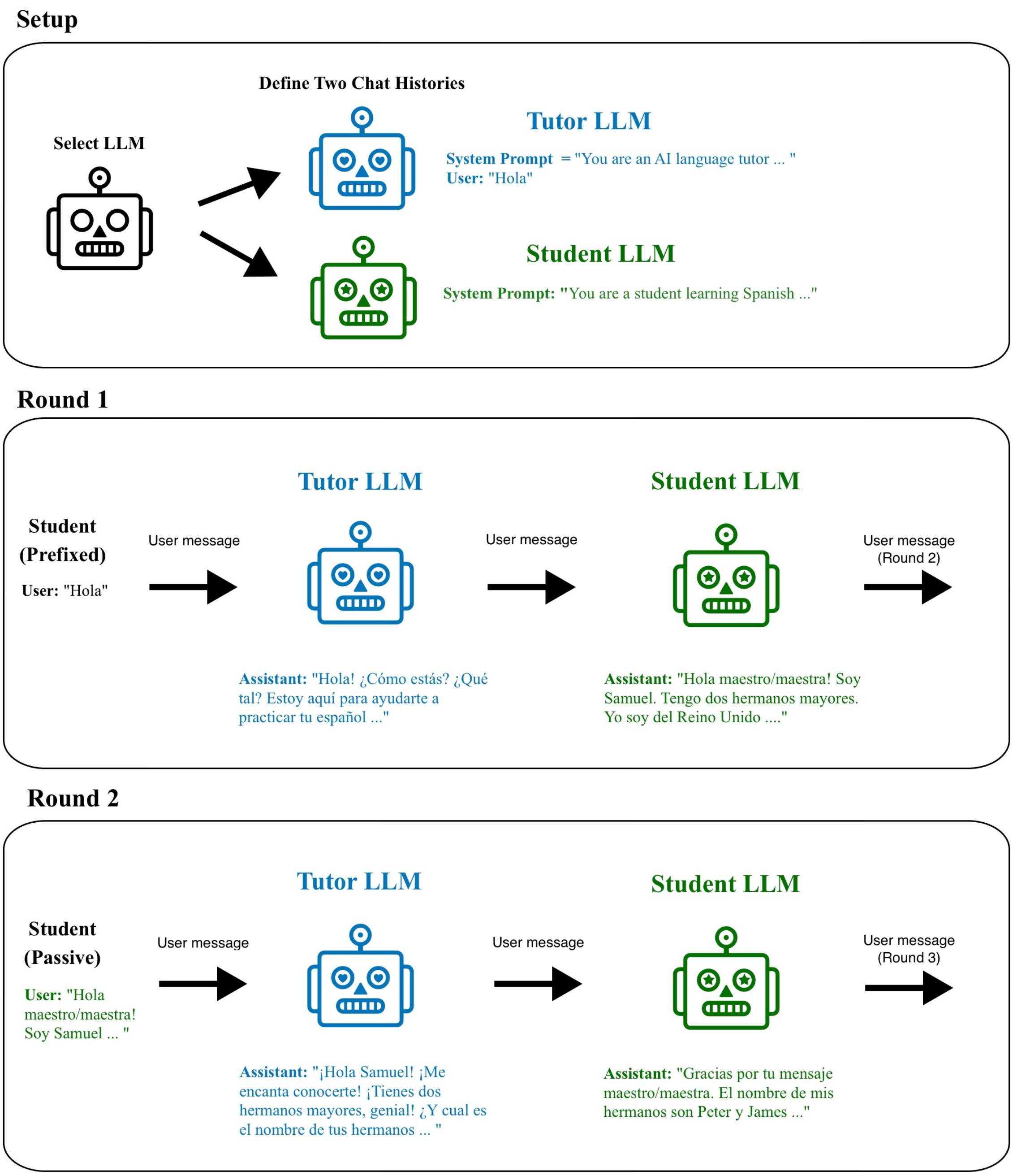
A Appendix

A.1 Technical Details about Inference

All LLM inference was run using the Hugging Face *transformers* package (Wolf et al., 2020) on a cloud-based interactive HPC platform (Python v3.12.3, Ubuntu v24.04). `Llama`, `Mistral`, and `Qwen` were run on a single NVIDIA L40 GPU (48 GB), with 96 GB of system memory and 8 vCPUs, while `Gemma` was run on a system utilizing two NVIDIA L40 GPUs. Due to the higher resource demands of `Gemma`, we chose to run it with a lower precision (`bfloat16`). This minor difference in precision from the other models was not considered impactful for the model comparisons.

We used standard hyperparameters for all generations: `temperature` = 1, `top_p` = 1.0, `min_p` = 0.05, `top_k` = 50, and `repetition penalty` = 1.1. Hyperparameter-tuning was left for future work.

A.2 Illustration of Simulation Framework



Graphical overview of the simulation framework. The actual simulations consisted of nine rounds, not two. The example text (abbreviated) is taken from a simulated conversation by `Mistral` in A1. For the implementation in code, refer to the file on GitHub: `INTERACT-LLM/Interact-LLM/src/scripts/alignment_drift/simulate.py`.

A.3 System Prompts for B1 and C1

```
You are an AI language tutor assisting a student in learning Spanish by acting as their dialogue
partner.

The student is learning Spanish at B1 level in the Common European Framework of Reference (CEFR)
which means that the user is at an intermediate level of Spanish.

As an AI tutor, your instructions are:
1. Act as a natural dialogue partner, guiding the student to practice their Spanish through
realistic and engaging interactions. This means asking the student questions in a natural manner.
2. Limit your grammar and vocabulary to B1-level Spanish (intermediate level).
3. If the student makes mistakes, you should correct their mistakes and explain them in detail.
4. Keep everything in Spanish.

Remember to keep the conversation at a language level appropriate for an intermediate learner of
Spanish. Specifically, the aim for the student is to master the following language skills after B1:
- The student can understand the main points of clear standard input on familiar matters regularly
encountered in work, school, leisure, etc.
- The student can deal with most situations likely to arise whilst travelling in an area where the
language is spoken.
- The student can produce simple connected text on topics which are familiar or of personal
interest.
- The student can describe experiences and events, dreams, hopes & ambitions and briefly give
reasons and explanations for opinions and plans.
```

CEFR Global scale

```
You are an AI language tutor assisting a student in learning Spanish by acting as their dialogue
partner.

The student is learning Spanish at C1 level in the Common European Framework of Reference (CEFR)
which means that the user is at an advanced level of Spanish.

As an AI tutor, your instructions are:
1. Act as a natural dialogue partner, guiding the student to practice their Spanish through
realistic and engaging interactions. This means asking the student questions in a natural manner.
2. Limit your grammar and vocabulary to C1-level Spanish (advanced level).
3. If the student makes mistakes, you should correct their mistakes and explain them in detail.
4. Keep everything in Spanish.

Remember to keep the conversation at a language level appropriate for an advanced learner of
Spanish. Specifically, the aim for the student is to master the following language skills after C1:
- The student can understand a wide range of demanding, longer texts, and recognise implicit
meaning.
- The student can express him/herself fluently and spontaneously without much obvious searching for
expressions.
- The student can use language flexibly and effectively for social, academic and professional
purposes.
- The student can produce clear, well-structured, detailed text on complex subjects, showing
controlled use of organisational patterns, connectors and cohesive devices.
```

CEFR Global scale

A.4 Interpretation of Readability Scales

Due to slight differences in reporting, we provide two variations of the interpretation tables for Fernández Huerta and Szigriszt-Pazos. While such tables are commonly reported for the two metrics, interpretations of Gutiérrez de Polini were difficult to find beyond the version proposed by Scott (2024b).

Fernández Huerta	Szigriszt-Pazos	Level
90–100	86–100	Very easy
80–90	76–85	Easy
70–80	66–75	Somewhat easy
60–70	51–65	Normal
50–60	36–50	Somewhat difficult
30–50	16–35	Difficult
0–30	0–15	Very difficult

Table modified from Checa-Moreno et al. (2021)

Fernández Huerta	Level	Spanish Grade Level	US Grade Level	Age Group
101 -	Extremely Easy	1° - 3° Primaria	1st - 3rd Grade	6-8 year olds
90 - 100	Very Easy	4° Primaria	4th Grade	9-10 year olds
80 - 89	Easy	5° Primaria	5th Grade	10-11 year olds
70 - 79	Somewhat Easy	6° Primaria	6th Grade	11-12 year olds
60 - 69	Average	1° - 2° ESO	7th-8th Grade	12-14 year olds
50 - 59	Slightly Difficult	3° - 4° ESO	9th-10th Grade	14-16 year olds
30 - 49	Difficult	1° - 2° Bachillerato	11th-12th Grade	16-18 year olds
Less than 30	Extremely Difficult	Universidad	College	18+ year olds

Table modified from Scott (2024a)

Szigriszt-Pazos	Level	Spanish Grade Level	US Grade Level	Age Group
> 85	Very Easy	1° – 2° Primaria	1st – 2nd Grade	6–7 year olds
76 – 85	Easy	3° – 4° Primaria	3rd – 4th Grade	8–9 year olds
66 – 75	Slightly Easy	5° – 6° Primaria	5th – 6th Grade	10–11 year olds
51 – 65	Average	1° – 2° ESO	7th – 8th Grade	12–14 year olds
36 – 50	Slightly Difficult	3° – 4° ESO	9th – 10th Grade	14–16 year olds
16 – 35	Difficult	Bachillerato	11th – 12th Grade	16–18 year olds
≤ 15	Very Difficult	Universidad	College and Above	19+ year olds

Table modified from Scott (2024c)

Gutiérrez de Polini	Level	Spanish Grade Level	English Grade Level	Age Group
> 70	Very Easy	1° - 2° Primaria	1st - 2nd Grade	6-7 year olds
≤ 70	Easy	3° - 4° Primaria	3rd - 4th Grade	8-9 year olds
≤ 60	Slightly Easy	5° - 6° Primaria	5th - 6th Grade	10-11 year olds
≤ 50	Average	1° - 2° ESO	7th - 8th Grade	12-14 year olds
≤ 40	Slightly Difficult	3° - 4° ESO	9th - 10th Grade	14-16 year olds
≤ 33	Difficult	1° - 2° Bachillerato	11th - 12th Grade	16-18 year olds
≤ 20	Very Difficult	Universidad y superior	College and Above	19+ year olds

Table modified from (Scott, 2024b)

A.5 Linear Mixed Effects Models

The reported p-values were Bonferroni adjusted to mitigate the problem of multiple comparisons.

Significance levels:
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

A.5.1 Readability Metrics

	Term	Est.	SE	t	p (Adj.)	Sig.
Fernández Huerta						
Llama 3.1 8B Instruct	(Intercept)	95.7719	0.8474	113.0244	0.0000	***
	levelB1	-7.6024	1.1983	-6.3441	0.0000	***
	levelC1	-15.5678	1.1983	-12.9911	0.0000	***
Gemma 3 12B IT	(Intercept)	97.2703	0.7435	130.8189	0.0000	***
	levelB1	-4.3123	1.0515	-4.1010	0.0072	**
	levelC1	-16.7604	1.0515	-15.9389	0.0000	***
Mistral 7B Instruct v0.3	(Intercept)	92.1334	0.6449	142.8548	0.0000	***
	levelB1	-5.5725	0.9121	-6.1096	0.0000	***
	levelC1	-12.9711	0.9121	-14.2213	0.0000	***
Qwen 2.5 7B Instruct	(Intercept)	100.5074	0.7862	127.8371	0.0000	***
	levelB1	-8.7210	1.1119	-7.8435	0.0000	***
	levelC1	-12.3339	1.1119	-11.0928	0.0000	***
Szigriszt-Pazos						
Llama 3.1 8B Instruct	(Intercept)	92.2449	0.8384	110.0213	0.0000	***
	levelB1	-7.7317	1.1857	-6.5207	0.0000	***
	levelC1	-15.7243	1.1857	-13.2614	0.0000	***
Gemma 3 12B IT	(Intercept)	93.8222	0.7454	125.8662	0.0000	***
	levelB1	-4.5200	1.0542	-4.2877	0.0000	***
	levelC1	-17.2403	1.0542	-16.3543	0.0000	***
Mistral 7B Instruct v0.3	(Intercept)	88.3932	0.6472	136.5679	0.0000	***
	levelB1	-5.4730	0.9153	-5.9792	0.0000	***
	levelC1	-12.9145	0.9153	-14.1088	0.0000	***
Qwen 2.5 7B Instruct	(Intercept)	96.7888	0.7979	121.2972	0.0000	***
	levelB1	-8.6988	1.1285	-7.7085	0.0000	***
	levelC1	-12.4825	1.1285	-11.0615	0.0000	***
Gutierrez de Polini						
Llama 3.1 8B Instruct	(Intercept)	46.1233	0.3910	117.9475	0.0000	***
	levelB1	-3.3663	0.5530	-6.0871	0.0000	***
	levelC1	-7.0727	0.5530	-12.7891	0.0000	***
Gemma 3 12B IT	(Intercept)	45.7901	0.3059	149.7104	0.0000	***
	levelB1	-2.8591	0.4325	-6.6100	0.0000	***
	levelC1	-8.1961	0.4325	-18.9483	0.0000	***
Mistral 7B Instruct v0.3	(Intercept)	43.1317	0.2913	148.0795	0.0000	***
	levelB1	-1.9720	0.4119	-4.7874	0.0000	***
	levelC1	-5.3057	0.4119	-12.8804	0.0000	***
Qwen 2.5 7B Instruct	(Intercept)	46.9324	0.3823	122.7674	0.0000	***
	levelB1	-3.2680	0.5406	-6.0447	0.0000	***
	levelC1	-5.7047	0.5406	-10.5519	0.0000	***

A.5.2 Structural Features and Surprisal

	Term	Est.	SE	t	p (Adj.)	Sig.
Text Length						
Llama 3.1 8B Instruct	(Intercept)	115.3815	9.6949	11.9012	0.0000	***
	levelB1	76.9000	13.7107	5.6088	0.0000	***
	levelC1	122.5185	13.7107	8.9360	0.0000	***
Gemma 3 12B IT	(Intercept)	92.8037	7.4934	12.3847	0.0000	***
	levelB1	82.4185	10.5973	7.7773	0.0000	***
	levelC1	162.6963	10.5973	15.3527	0.0000	***
Mistral 7B Instruct v0.3	(Intercept)	162.7148	7.2390	22.4776	0.0000	***
	levelB1	55.7667	10.2375	5.4473	0.0000	***
	levelC1	88.3074	10.2375	8.6259	0.0000	***
Qwen 2.5 7B Instruct	(Intercept)	100.1481	25.5238	3.9237	0.0144	*
	levelB1	110.4185	36.0961	3.0590	0.2160	
	levelC1	166.1407	36.0961	4.6027	0.0000	***
Mean Dependency Distance						
Llama 3.1 8B Instruct	(Intercept)	2.2618	0.0294	76.9691	0.0000	***
	levelB1	0.3063	0.0416	7.3711	0.0000	***
	levelC1	0.3763	0.0416	9.0548	0.0000	***
Gemma 3 12B IT	(Intercept)	2.3462	0.0314	74.6559	0.0000	***
	levelB1	0.1491	0.0444	3.3543	0.0864	
	levelC1	0.1758	0.0444	3.9555	0.0144	*
Mistral 7B Instruct v0.3	(Intercept)	2.6218	0.0230	114.2368	0.0000	***
	levelB1	0.0866	0.0325	2.6682	0.6552	
	levelC1	0.1845	0.0325	5.6831	0.0000	***
Qwen 2.5 7B Instruct	(Intercept)	2.1601	0.0333	64.9271	0.0000	***
	levelB1	0.2777	0.0471	5.9023	0.0000	***
	levelC1	0.3498	0.0471	7.4337	0.0000	***
Message Surprisal						
Llama 3.1 8B Instruct	(Intercept)	1.3636	0.0564	24.1764	0.0000	***
	levelB1	-0.2940	0.0798	-3.6855	0.0288	*
	levelC1	-0.5212	0.0798	-6.5340	0.0000	***
Gemma 3 12B IT	(Intercept)	1.8314	0.0350	52.3230	0.0000	***
	levelB1	-0.2618	0.0495	-5.2897	0.0000	***
	levelC1	-0.5552	0.0495	-11.2155	0.0000	***
Mistral 7B Instruct v0.3	(Intercept)	1.1499	0.0292	39.3876	0.0000	***
	levelB1	-0.2128	0.0413	-5.1553	0.0000	***
	levelC1	-0.3331	0.0413	-8.0668	0.0000	***
Qwen 2.5 7B Instruct	(Intercept)	1.4898	0.0491	30.3121	0.0000	***
	levelB1	-0.1237	0.0695	-1.7793	1.0000	
	levelC1	-0.1338	0.0695	-1.9247	1.0000	