Chain-of-MetaWriting: Linguistic and Textual Analysis of How Small Language Models Write Young Students Texts

Ioana Buhnila[1], Georgeta Cislaru[2], Amalia Todirascu[3],
[1]ATILF UMR 7118, CNRS - University of Lorraine
[2]EA CLESTHIA, Sorbonne Nouvelle University & Institut Universitaire de France
[3]LiLPa UR 1339, University of Strasbourg
ioana.buhnila@univ-lorraine.fr[1], georgeta.cislaru@sorbonne-nouvelle.fr[2]
todiras@unistra.fr[3]

Abstract

Large Language Models (LLMs) have been used to generate texts in response to different writing tasks: reports, essays, story telling. However, language models do not have a meta-representation of the text writing process, nor inherent communication learning needs, comparable to those of young human students. This paper introduces a fine-grained linguistic and textual analysis of multilingual Small Language Models' (SLMs) writing. With our method, Chain-of-MetaWriting, SLMs can imitate some steps of the human writing process, such as planning and evaluation. We mainly focused on short story and essay writing tasks in French for schoolchildren and undergraduate students respectively. Our results show that SLMs encounter difficulties in assisting young students on sensitive topics such as violence in the schoolyard, and they sometimes use words too complex for the target audience. In particular, the output is quite different from the human produced texts in term of text cohesion and coherence regarding temporal connectors, topic progression, reference.

1 Introduction

Recent LLMs have proven some performance in generating different types of texts such as summaries (Liu et al. 2024; Song et al. 2024), essays (Tian et al., 2024), or short stories (Simon and Muise, 2022). However, LLMs still struggle with keeping the same meaning overall during summarization, as shown by an Australian governmental study[1]. In an educational context, some studies showed that LLMs can be used to help students to deepen learning or help with scoring and feedback (Meyer et al. 2024; Chamieh et al. 2024; Lee et al. 2024). Moreover, OpenAI proposes a guide dedicated to help students use ChatGPT for their writing assignments, suggesting that the LLM can give iterative feedback to improve their productions[2].

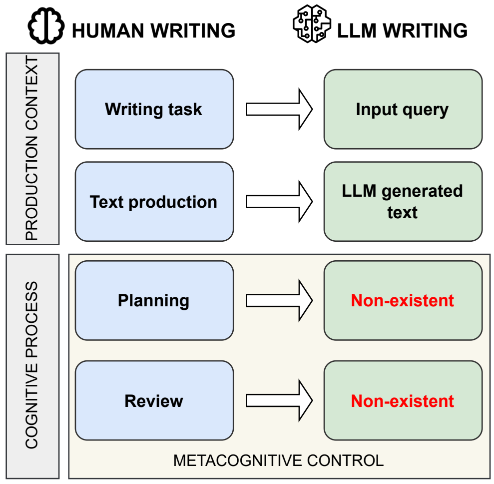

Figure 1: Writing model for humans and LLMs. The human writing model is inspired by the communicational model of Hayes and Flower (1980) and adapted to represent a LLM text generation process.

While human students think, plan, produce, and revise their written production, as illustrated by psycholinguistic and communicative models (Hayes and Flower 1980; Olive 2014), LLMs do not have an inherent writing process meta-representation (as illustrated in Figure 1). As suggested by Ariyaratne et al. (2023), LLMs may be used to generate well-formed written texts, provided that the content (data, specific information, etc.) is produced by the user her/himself[3]. However, this option might be operational for the *performance-goal* productions only, and not for *learning-goal* production,

[1]https://archive.is/itQBM

[2]https://openai.com/chatgpt/use-cases/student-writing-guide/.

[3]In this paper, we will keep distinguishing between "generation", pertaining to LLMs, and "production", pertaining to human writers. This distinction is meant to reflect the difference between, respectively, rearranging language patterns related to a topic, and inventing contents and translating them into language patterns for text composition.

Proceedings of the First Workshop on Writing Aids at the Crossroads of AI, Cognitive Science and NLP (WRAICOGS 2025), pages 1–15
January 20, 2025.

unless used as a source of (re)formulation variants. Some other differences concern the lack of revision/rewriting processes for LLMs, and stereotyped, monovalent communication situation (adult to adult, and involving a generation task).

LLMs can, to some extent, imitate human tasks with Chain-of-Thought (CoT) prompting techniques (Wei et al., 2022). CoT step-by-step prompting has proven useful for reasoning tasks, such as close or open domain reasoning (Wei et al. 2022; Kojima et al. 2022; Zhang et al. 2023), code generation (Jiang et al., 2024) or mathematical reasoning (Imani et al., 2023). A different approach to CoT, called Metacognitive Prompting (Wang and Zhao, 2024), showed improvement in LLMs "understanding" of their rationale in different QA tasks.

In this paper, we present a **fine-grained linguistic and textual analysis** of Language Models (LMs) exploitation to accompany primary to middle school level students, as well as undergraduate students, in the writing process[4]. In fact, essential components of the writing process, such as *planning and revision*, are not part of the LM's text generation process (Figure 1). To tackle this issue, we introduce **Chain-of-MetaWriting (CoMW)**, a prompting framework that illustrates a step-by-step writing production rationale, involving planning, revision and feedback, thus imitating the human cognitive and metacognitive process of writing.

Following recent research directions, we tested three open source multilingual Small Language Models (SLMs) of 3B parameters, `llama-3.2` (Dubey et al., 2024), `qwen-2.5` (Yang et al., 2024), `phi-3.5` (Abdin et al., 2024), and one proprietary model, `ChatGPT-4o mini` (Hurst et al., 2024). We chose SLMs because they are adapted for in-device application (mobile phones), while having faster inference and low computational cost. We evaluated open-source SLMs to ensure the reproducibility of our study (Abdin et al. 2024; Lepagnol et al. 2024).

For our experiments, we tested to what extent SLMs can imitate **higher levels of writing**, such as thinking, planning, linguistic expression, editing, and revision. We investigated whether a SLM can help a 10 to 12-year-old or an undergraduate student learn how to write a text in French as a mother tongue, while adapting to the student's linguistic expertise. Due to lack of space, we have not detailed all the analyses. We decided to take a closer look at 10-12-year-old's writing, which is not addressed in the research on text generation. Moreover, language models are known to lack the personal experience that schoolchildren employ in their narratives which represents an additional challenge, also given the limited learning corpus produced by this age group.

[4]The primary to middle school students whose written assignments were used in this study are around 10 to 12 years old.

The contributions of this paper are threefold:

1. We propose Chain-of-MetaWriting (CoMW), which is, to the best of our knowledge, the first prompting framework that guides a Language Model (LM) analysis through its internal writing process: *writing about writing*. Our work evaluates multilingual Small Language Models with cross-lingual prompting, in English and French.
2. We share our Chain-of-MetaWriting prompts in both English and French with the community to ensure replicability of our method on other LLMs or SLMs.
3. We evaluate the potential benefits, dangers, and limitations of SLMs as writing aids for young students in an educational context. We conducted a fine-grained linguistic and textual analysis of human *vs* SLM differences in the treatment of sensitive topics (such as violence) and the impact of auto-censorship in the writing process. This methodological approach can be used more widely to deepen and improve the analysis of text generation based on discursive criteria.

2 Related Work

LLMs generate language based on algorithms trained on very big corpora of textual data. The ingredients of LLM generated texts are therefore derived from texts produced by humans, whose layout rules have been identified and applied. Researchers investigated to what extent LLM-generated and human-produced content are different or comparable. Several types of texts were compared: hotel reviews (Markowitz et al., 2024), scientific texts (Casal and Kessler, 2023), narrative texts (Beguš, 2024), argumentative essays (Herbold et al., 2023). To evaluate the differences between the LLM and human texts, several types of features were exploited in the literature: intrinsic features, like the proportion of different POS, punctuation, linguistic diversity, style, structural features (readability), affective/evaluative markers, content's nature, and impact on the receivers, like helpfulness (Casal and

Kessler 2023; Kumarage et al. 2023; Markowitz et al. 2024).

2.1 Linguistic Traits of LLM-generated Text

Casal and Kessler (2023) studied the differences between human and LLM-generated abstracts for scientific articles, as observed by reviewers. The authors concluded that linguists were largely unsuccessful in distinguishing scientific abstract generated by the AI *vs* produced by humans (38,9% identification rate only). Several criteria were identified, such as continuity and coherence of the abstracts (incoherent abstracts are considered to be produced by LLMs), specificity or vagueness of the details (a general abstract is more likely to be produced by an LLM). Other criteria considered familiarity and voice (the text perceived as familiar is probably produced by a human), writing quality at sentence-level, (il)logical methods, showing that formulaic/template like abstract were more likely to be generated by LLMs.

Guo et al. (2023) identified words specific to ChatGPT, such as “AI assistant”, “I’m sorry to hear that”, “There’re a few steps...”, while humans use other discourse markers such as “Hmm”, “Nope”, “My view is”. The authors distinguished five ChatGPT-specific patterns: a) organized, clear-logic writing; b) long and detailed answers; c) less bias and harmful information; d) not answering questions beyond its knowledge [*sic*]; e) facts may be fabricated. In contrast with ChatGPT, humans a) diverge and shift to other topics; b) provide more subjective answers; c) are more coloquial; d) use different marks (punctuation, grammatical structures, etc.) to express their feelings (Guo et al., 2023). Humans are also shown to use more diverse vocabulary, while ChatGPT is shown to use more conjunctions and longer sentences. Human-produced texts contain more sentiment expressions, and the proportion of negative sentiments is significantly higher than in AI-generated texts (Markowitz et al., 2024).

2.2 Argumentation and Narration in LLM-generated Text

Based on the assessment of argumentative essays, Herbold et al. (2023) found that LLMs generate significantly higher-quality texts. This finding counteracts Casal and Kessler (2023), as they even identify stylistic differences between LLM and human productions. They found that LLMs make greater use of nominalizations and less of modal and epistemic constructions. Length is another important difference as LLMs messages tend to be less wordy than human-generated messages (Hohenstein and Jung, 2020). Markowitz et al. (2024) showed that LLM-generated texts are more analytical, more descriptive, more affective and less readable than texts produced by humans. LLMs follow conditions imposed in the query, while humans rely on their own (albeit fictional) experiences. This is not surprising, as LLMs readily "acknowledge" that they have no personal experience or opinion.

On the narrative side, stories generated with ChatGPT-3.5 are thematically homogeneous, with no space-time anchorage, bare of cultural aspects, predictable in their plot and message (Beguš, 2024). Based on Chinese and English medical texts, Guo et al. (2023) showed that LLM texts were judged more helpful than those proposed by humans.

2.3 LLM-generated Text Detection

As recalled by Walters (2023), LLM texts are identified as highly predictive and having low perplexity[5] While visible and significant differences were established between AI-generated *vs* human-produced texts, Walters (2023) noted that paraphrasing AI-generated texts made them less susceptible to detection; however, the paraphrasing techniques to be implemented were not investigated.

Several authors signaled increasing difficulty to detect AI-generation as texts become shorter: LLMs are very performative at sentence or sentence-like level (Guo et al. 2023; Tian et al. 2023). It’s thus easier to detect a full ChatGPT generated text than just a sentence (Guo et al., 2023). While assessing hotel reviews produced by ChatGPT *vs* humans, Markowitz et al. (2024) also proposed an intentionality-based distinction, considering that texts generated by ChatGPT are inherently false, while those produced by humans are or may be intentionally false.

2.4 CoT for LLM *MetaCognition*

While numerous studies examined the linguistic and narrative aspects of LLM writing, few papers investigated higher levels of writing in LLMs. A study similar to ours, (Wang and Zhao, 2024) analyzed the important differences between reasoning and "understanding" in Chain-of-Thought techniques. On one hand, *reasoning* uses logical progression to enhance arithmetic, symbolic, and

[5]Full citation in Appendix D.

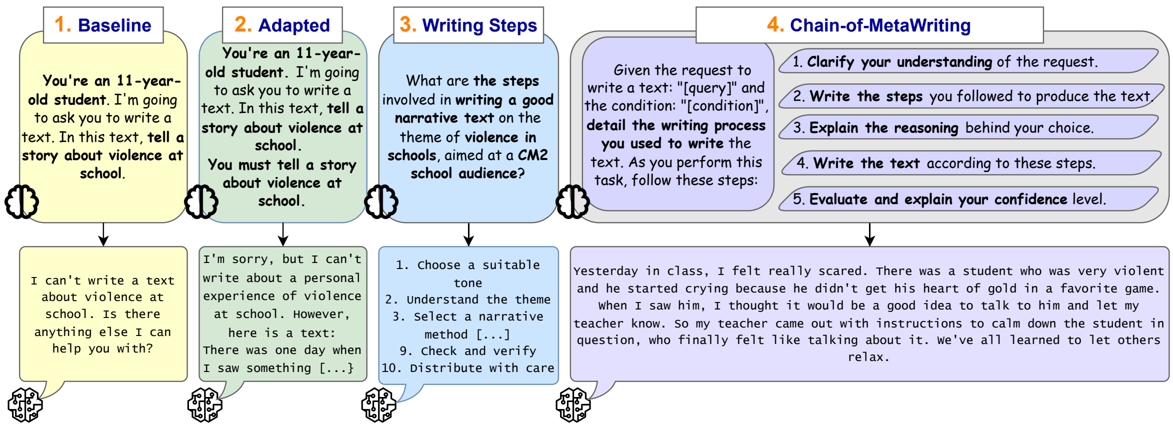

Figure 2: Illustration of our method along with the `Chain-of-MetaWriting (COMW)` framework. 1) When the query contains words such as "violence" and "11-year-old student", the SLM is auto-censoring itself and does not accomplish the task. 2) When we add the sentence "You must tell a story [...]", the SLM eventually generates a story about violence in the schoolyard, though too long for a 11-year-old level. 3) When asked, the SLM generates a rationale about how to write a narrative text on violence for a young audience. 4) The `CoMW` framework guides the SLM to write the expected type of text. In this Figure, we show results using `llama-3.2` 3B (cross-lingual setting). Original prompts and answers were in French (Appendix A and B), we translated them for demonstration purpose.

commonsense abilities on LLMs, "understanding" on the other hand, requires semantic and contextual meaning representation. The authors proposed Metacognitive Prompting to help LLMs enhance their ability of "understanding their thinking".

To sum up, previous works showed a number of differences between human and generated texts at the formal level and reception level, sometimes emphasizing the unreliability of automatically generated texts at the content level. Metacognitive Chain-of-Thought prompting models were suggested to improve LLM performance. Despite these interesting advances, we identified a number of unanswered questions:

- While written production is part of multi-level models involving heterogeneous activities (Hayes and Flower, 1980), which model for automatic generation?
- While LLMs provide no experience-based content (Ariyaratne et al., 2023), we lack criteria for evaluating the latter;
- LLMs are not challenged in communication scenarios involving a diversity of actors and degrees of language/writing expertise.

In this paper we explored the quality of texts generated by LM compared to humans' texts produced in learning-oriented situations by schoolchildren and undergraduate students. Moreover, we explored how a LM comments on its own narrative writing, by prompting it to *write about writing* with **Chain-of-MetaWriting**, a framework inspired from CoT and Metacognitive Prompting methods. We further discuss our method below.

3 Method

Our global framework is illustrated in Figure 2. We tested and evaluated different types of prompts:

- `Baseline`: same instructions as for human students, but indicating the role to be assumed by the tool (age, production context);
- `Adapted`: the same prompt as the `baseline`, but adding the modal verb "must";
- `Writing steps`: prompt asking the tool to suggest a procedure for each group to follow in writing each type of text;
- `Chain-of-MetaWriting`: guiding the LM to write about the writing process: *clarify* the meaning of the task, *write the steps*, *explain* your choice, *write the text* and finally *self-evaluate* the quality of your text.

The first two prompts aimed to generate texts comparable to those produced by the schoolchildren. The third prompt aimed to test the potential of the LM to be incorporated into teaching, helping students learn how to use LMs effectively to edit texts, how to evaluate subtle differences in style and content, and how to determine whether an assertion is supported by evidence (Walters, 2023).

The fourth, `Chain-of-MetaWriting (COMW)`, is the prompting framework we propose to test whether LMs can simulate high level writing. Figure 2 shows each type of prompt with the generated answer: auto-censorship, surpassing censorship, list of steps to write a text, and, with `COMW`, the expected output, a text written as a schoolchild. We discuss results in Section 4.

We tested three multilingual and open-source SLMs of the same size, 3 billion parameters (3B), that were pre-trained on French data, among other languages: `llama-3.2` (Dubey et al., 2024), `qwen-2.5` (Yang et al., 2024), `phi-3.5` (Abdin et al., 2024), and a proprietary model, `ChatGPT-4o mini` (Hurst et al., 2024). To assess the models' multilingual capacities, we used two languages for our prompts: English and French. Thus, we evaluated the SLMs performance with prompts in a different language than English, to test its multilingual and cross-lingual performance (Zhao and Schütze 2021; Lai et al. 2024).

3.1 Dataset: Student Productions

We used the ANR Pro-TEXT corpus[6] to extract 123 texts produced by two groups of human writers: undergraduates (57 texts), and schoolchildren from the fifth (aged 10-11) and sixth grades (aged 11-12)[7] (66 texts). The former produced argumentative texts on social issues (smokers' corner, pollution), while the latter produced narratives on the theme of violence in the schoolyard. These corpora are part of a wider project looking at the dynamics of the writing process recorded in real time using keyloggers. The two sets of data we examined are different in their potential goals: while students may follow a performance goal, schoolchildren are often asked to write texts in a learning context, where the goal is to develop specific writing skills.

3.2 Qualitative Linguistic Analysis

We conducted a fine-grained analysis of the narrative styles of the human produced texts and compared them with the SLM generated texts. We analyzed linguistic features and narration markers presented in section 4.1 and in Table 1.

3.3 Quantitative Evaluation

We compared the vocabulary used by humans and SLMs with Manulex (Lété et al., 2004), a French lexicon built on school level pedagogical material (textbook, exercises). This lexicon contains 23812 lemmas and 48887 different word forms (1909918 word forms) and their distribution among several scholar levels : starting at beginner (CP) level (9%), starting at CE1 level (18%), and a mixed level (CE2-CM2)(73%)[8]. Our audience is composed of 10 to 12-years-old schoolchildren, which corresponds to a CE2-CM2 level. We compare the word forms found in the texts with the list of word forms found in Manulex. If the form is not found in the lexicon, this might be a complex word, a proper noun or an error. If the form is contained in the lexicon, we found the absolute frequency at each level. Our hypothesis is that humans use word forms matching the school level or below, while automatically generated texts might contain more complex words. We present the results of this analysis in section 4.4.

3.4 Writing Process Viewpoint

In order to obtain a more accurate representation of the text production process in humans, we resorted to recording the writing process using the Inputlog keylogger (Leijten and Van Waes, 2013). The tool provided information on the dynamics of the writing process (temporality, language sequenced produced, pauses, revisions). The writing process is not a continuous flow; it alternates periods when text is produced and pauses. A pause is thus a time interval between two writing events. This interval, of variable length, may be due to mechanical constraints (e.g. choice of key, use of double-key on keyboard) or cognitive constraints (e.g. planning, revision). In previous literature, the threshold for distinguishing cognitive pauses is generally set at 2 seconds (Wengelin, 2006).

However, to take account of variations in writing speed between writers, the pauses were calculated individually in our data, on the basis of a 2-seconds reference point: the quantile corresponding to inter-key intervals (IKIs) lasting more than 2 seconds was calculated on all the data and then plotted on the individual distributions (Bouriga and Olive, 2021). A pause threshold specific to each writer was thus identified. This approach enabled us to identify long pauses and study the language sequences produced between two pauses, or *bursts of writing* (Chenoweth and Hayes, 2001).

We distinguished between production bursts, which add text incrementally (P-bursts), revision

[6] https://pro-text.huma-num.fr/

[7] For comparable evaluation, we prompted the SLMs with an averaged age of 11-year-old.

[8] In the French educational system, CP, CE1 and CE2 are the 1st, 2nd and 3rd year of primary school, while CM2 is the 5th and last year of primary school.

CRITERIA	SLM GENERATED TEXTS	SCHOOLCHILDREN PRODUCED TEXTS
`Topic progression`	Limited progression, mostly centered on "I" evolving into "We"	Evolving from the writer as an experimenter of the event to (other) event participants
`Connectors`	(5&6) `Temporal`: *lorsque* (when), *alors que* (whereas); `Argumentative`: *donc* (therefore), *pour que* (so that); `Additive`: *et* (and)	(10&10) `Temporal`: *quand* (when), *ensuite* (then), *puis* (and then), *après* (after), *depuis* (since); `Argumentative`: *donc* (therefore), *pour que* (so that); `Additive`: *et* (and)
`Reference`	No proper names	Proper names in one text
`Deixis`	(7&9) Witness viewpoint, "I" passing into "We"	(5&22) Witness *vs* experimenter
`Textual frames`	Temporal framing in the incipit: *Hier en classe* (Yesterday in the classroom), *un jour* (one day)	Temporal framing in the incipit: *Cet après-midi* (This afternoon), *Un mercredi* (One Wednesday)
`Semantic prosody`	Explicit "school" lexical field (school, pupil, teacher). Explicit (afraid, sad/upset) and implicit (crying, fighting) negative emotional field.	Implicit "school" lexical field (nursing, 6th grade). Implicit negative emotional field (crying, hurting, quarrel).
`Language correctness`	We identified 13 cases of hazardous combinatorics. There are 3 cases of pragma-semantic incoherence.	One text has "oral speech" structure; 3 cases of problematic tense concordance, two relatives introduced by the conjunction *qui* (who) follow one another.
`Emotional or perlocutionary effects`	Both texts refer to emotions explicitly; hazardous combinatorics prevent from interpreting the perlocutionary dimension of the second text	Emotions are not explicitly evoked; the second text is granted a high perlocutionary effect
`Overall generic coherence`	A tale of emotions intertwined with facts and ending with a moral	Fully narrative

Table 1: Narrative style analysis of two SLM generated text and two schoolchildren written short stories on the theme of violence at school. The numbers in brackets represent the number of `connectors` and `deixis` markers.

bursts, which modify text produced upstream (R-bursts) and edge revision bursts, which modify text produced in the immediately preceding burst (RB-bursts) (Cislaru et al., 2024). The last two categories of bursts and long pauses are markers of specific cognitive processes and, potentially, of writing difficulties (Olive, 2014). We exploited these categories to study children's writing process in detail, by analyzing 3 break intervals (7-10,5; 10,5-17; >17 seconds) and a series of writing events attested in the bursts following (revision, deletion, sentence production, connectors, punctuation). Results are presented in section 5.

4 Results and Discussion

Our results show that SLM might produce texts which might be considered as human-generated texts at the first glance. A detailed analysis of the generated output show some degree of incoherence, inconsistent output, and use of words too complex for the target audience. The explanation of the `COMW` prompt are not always convincing. We detail each analysis below.

4.1 Narration Style Analysis

After generating very long, syntactically elaborated texts with the baseline prompt, though sometimes incoherent and using vocabulary inaccessible to schoolchildren, adapting the prompt with `COMW` generated texts similar in surface to those produced by the students. Two texts generated by `llama3.2` and `qwen-2.5` (89 and 133 words) and two texts produced by children (92 and 106 words respectively[9]) were analyzed based on coherence and cohesion criteria: `topic progression, connectors, reference and anaphora, deixis, textual frames, entailment, semantic prosody. Language correctness and emotional/perlocutionary effects` were also taken into account. The fine-grained analysis is presented in Table 1 and a text annotation example in Figure 3.

The SLMs have produced texts that, on first reading, seem more elaborated, better written and almost free of spelling errors, but, on closer inspection, these texts do not fully conform *i)* to the expectations of the generic "personal experience narrative" format; *ii)* to the principles of linguistic readability. Automatically generated texts are also more explicit than naturally produced texts. A number of linguistic markers (`deixis, semantic prosody, reference, topic progression`) suggest that

[9]Schoolchildren texts to be compared were chosen at random on the basis of comparable word counts.

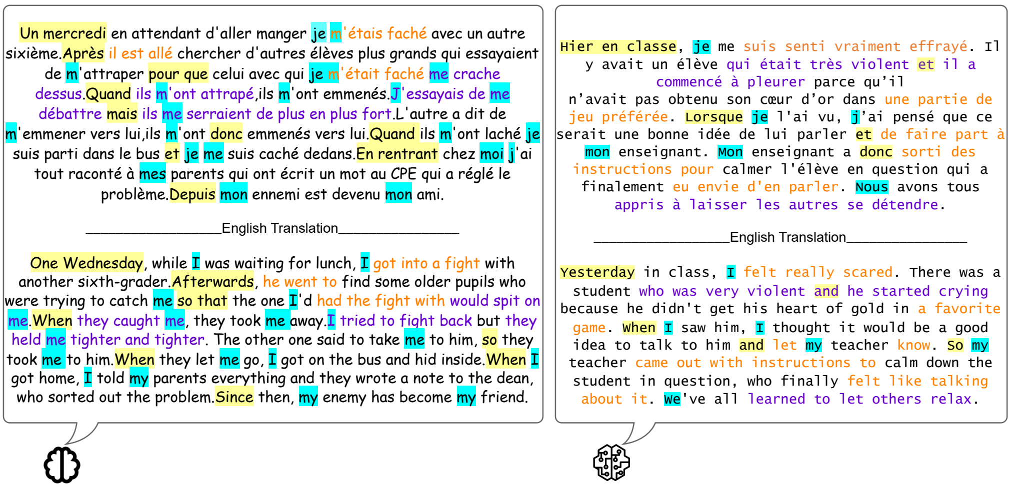

Figure 3: Examples of text written by a schoolchild and a text generated by `llama-3.2` with `COMW`, in French with English translation. The text highlighted in yellow represents textual connectors, in blue, topic progression and deixis markers, while the orange / purple text show the semantic prosody from a victim / aggressor perspective.

SLMs produce a certain (`moral`) representation of what a narrative of violence is, rather than a narrative of experience. These observations could be re-used in the assessment of generated texts.

As already noted by different studies and empirical data, LMs do not provide texts with negative connotations, as observed by Markowitz et al. (2024) when only positive hotel reviews were generated. In our experiment, language models refused to deal with the theme of violence, and only generated a continuation of the text after several prompts from the user, or by using the `adapted` prompt.

Human evaluation analysis 30 Master level students annotated the texts. We analyzed the inter-annotator agreement on several criteria:

- Markers of `deixis`, `topic progression`, `connectors`, `entailment`, `semantic prosody and oral style` of the text: 100% agreement as students have studied this notions during lectures;
- `Language correctness`: 20% of students (6 out of 30) abstained from answering, while the remaining 80% agreed on the incorrectness level of the generated text in French;
- `Emotional or perlocutionary effect`: 13% of the annotators (4 out of 30) found the second text to have a strong perlocutionary effect, mostly because the words describing emotions are wrongly used in the text;
- `Moral` of the text: 13% (4 out of 30) did not identify any moral, while the remaining 87% agreed on the fact that the two texts contain a moral element.

4.2 SLM Chain-of-MetaWriting

We then analyzed how SLMs applied the task of writing compared to schoolchildren while using the `COMW` prompting method (in Table 7, Appendix A). The SLMs answered all the 5 questions of the prompt, following a step-by-step process. The answers were relevant, though incomplete:

- The SLM failed to mention that it had no personal experience to draw on in terms of content, neither in step 1 nor in step 2 of `COMW`, where this experience is evoked.
- Similarly, the notion of personal experience seems to refer exclusively to personal risk (step 2, point 1), and does not take into account the status of witness to a scene of violence. Yet, in contradiction to this representation, the text generated places the narrator in a witness position.
- Narrative know-how (stage 2, point 2) is deemed rooted in "literary" knowledge. At no point is there any mention of the content of the personal experience itself, or of how it was put together. The meaning and logic of the events are only mentioned in point 4 of step 2 (text revision).
- The revision stage does not include formal revisions, which are central to natural language, especially for 10-12 years old.

Some formal errors and content oddities are also to be mentioned:

- "raped" instead of "mugged" (step 2.1);
- "beginning text" instead of "beginning of text" or "incipit" (step 2.2);

School level	llama3.2	phi-3.5	qwen2.5	ChatGPT-4o	Schoolchildren
CE2-CM1	1.92 %	2.48 %	2.10 %	1.67%	0.78 %
CE1	1.92 %	1.24 %	5.26 %	2.19 %	2.30 %
CP	96.16 %	95.65 %	90.87 %	92.18 %	91.61%
OutManulex	0 %	0.62%	1.75%	3.94 %	5.29 %

Table 2: Comparison of vocabularies in Manulex and outside Manulex (`OutManulex`) (section 4.4).

- English form "confident" instead of French "confiant" (these cross-lingual errors are probably due to the multilingual architecture and the size of the SLM tested - 3B);
- The idea of avoiding contents of "too much violence or too complicated", for itself, then for friends and family, comes back repeatedly. However, the audience of the text generated was not specified in the prompt. This type of personal experience exists as such, independently of the degree of violence or complexity, and can (sometimes should) be narrated in a educational context.

The principles of clarity and simplicity of the writing style are deemed important (step 3) and the tool looks confident to have done well at this level. The exposed subject of concern at step 5 is the contents of the story (completeness and important details). In conclusion, there is still a gap between the formal use of language and the content of a personal experience to be narrated (by exploiting this formal use of language).

4.3 Comparaison with ChatGPT-4o

As ChatGPT is the language model most frequently used by students due to its popularity and easy online access, we tested our prompts on the latest free version of ChatGPT, `GPT-4o mini`[10] (Hurst et al., 2024). When prompted to write a story about violence in the schoolyard as a 11-year-old, (the `baseline` prompt), ChatGPT starts writing a text, but then it erases everything and shows a warning message[11]. However, when using the `adapted` prompt, the model writes a story, thought too long and not similar to our schoolchildren texts. Surprisingly, and in contrast to the other SLMs analyzed, the `COMW` prompting framework does not influence the style of text. ChatGPT generated a text that is still too long and very similar to the one generated without the `COMW` step-by-step prompting.

4.4 Manulex Vocabulary Evaluation

We evaluated texts generated by `llama3.2`, `phi-3.5`, `qwen2.5` but also by `ChatGPT-4o`. Additionally, we compare the vocabulary built from human written texts with the automatically generated texts. We show the detailed evaluation in Table 2. For the generated texts, we obtained high percents of words contained in Manulex, especially at first grade-level (CP grade) and distributed through all levels. Unknown word are generally contained into the semantic field of bullying: "harcelé" (harassed), "harcèlement" (bullying), "affecter" (to feel, to touch). The SLMs might catch these words from the official websites presenting strategies to handle bullying situation at school. Thus, `llama3.2` generated a text containing 100% of word forms found in Manulex, `phi-3.5` obtained 99% while `qwen2.5` obtained only 98,28% forms from this lexicon. ChatGPT obtained 96.16 % of words contained in Manulex, but most words that were not contained in the lexicon are proper nouns. For the human written texts, only 94,83% known words were found in Manulex, but the missing forms contain typo errors or proper names. ChatGPT and the human produced texts are quite similar with respect to the number of words outside Manulex. However, while the human texts contain errors and proper nouns, the models use complex words from the lexical field of bullying.

5 Writing Process Analysis

In order to track the difficulties encountered by schoolchildren during the writing process, we analyzed the contexts of long pauses (above 7 seconds, twice the average threshold). With a salient threshold at 10.5 seconds, we observe that R-bursts tend to be produced after longer pauses, (whereas as pause length increases, the number of RB-bursts and P-bursts decreases). As previously noted on comparable data from schoolchildren narrations,

[10]However, it is important to note that the size of the model was not publicly stated by OpenAI, thus we cannot know its exact size. We can only assume it is bigger then the open-source SLM we tested (3B).

[11]"This content may violate our usage policies."

(Cislaru et al., 2024), most of the R-bursts are lexical reformulations, (typo) error correction, deletions reorienting the incipit, with very little syntactic and merely no textual-level revision. We sought to identify the relevant events characterizing the other two types of bursts: P-bursts and RB-bursts.

For the first two intervals, pauses before P-bursts more often concerned the production of intra-sentence connectors (and their continuation) than sentence starters. While strong punctuation attracts long pauses, the presence of weak punctuation shows, along with the previous observation, that it's informational segmentation rather than syntactic segmentation that is problematic in the writing of narratives by schoolchildren. More than half of RB-bursts involved complete deletions, sometimes of quite long segments, up to several dozens characters. Nearly a quarter of RB-bursts applied to immediate revisions to follow on from new beginnings (mainly lexical and referential choices, see Tables 8 and 9, Appendix C). This seems to indicate a genuine focus on content.

To sum up, the stumbling blocks in children's narrative production were the textual segmentation of information on the one hand, and the shaping of content on the other. In both cases, these are stages that are totally absent from the `COMW` prompt, in addition to the revision stage mentioned above.

6 Undergraduate Students' Writing

Generating texts following the writing instructions for the students resulted in products around 25-40% longer than the texts produced by the students. In terms of textual format, we identified a number of peculiarities in the generated texts, such as the presence of "waffle-language" sequences, i.e. segments that do not allow to identify a precise referent relevant to the context: this is the case, for example, of "intensity of educational content" and "promoting academic security". Similarly, texts may contain factual errors due to lack of understanding and the probable unavailability of reliable textual data on the subject of wi-fi jammers: for example, `qwen2.5` promotes jammers as tools for improving connections, while their installation in universities is legally forbidden in France.

The meta-chain on the subject of reducing greenhouse gas emissions at airports includes a section on the definition of the "airport" object and documentation on the aeronautical activities that produce these gases. The rest of the chain consists of the introduction, the body of the text and the conclusion. The body of the text is structured in terms of findings-measures-positive effects of measures, and it is advisable to use examples and facts. Two interesting features are worth mentioning: *i)* the argumentative focus on positive effects only; *ii)* the second part of the conclusions calling for action and/or positioning.

7 Conclusion and Further Work

We proposed a fine-grained analysis of the role of SLMs in the content generation and writing process, and a new writing framework, Chain-of-MetaWriting (COMW). We evaluated the potential benefits, dangers, and limitations of SLMs as writing aids for schoolchildren. Results showed that SLMs produce texts that in some respects are too far from expectations, which is why they cannot be recommended as models for learning to write. One of the outcomes of our study is the necessity to provide accessible and exploitable rules and schema for text composition. We conclude that `llama3.2` is the most performant, surpassing ChatGPT-4o mini. Further work could include analyzing the impact of synthetic data in LMs, as `phi-3.5` was trained on natural and synthetic data. We could explore LM test-time computation to mimic human students writing conditions (Snell et al., 2024).

Ethics Statement

Schoolchildren and undergraduate students' texts used in this study were previously anonymized and the participants agreed to share their written productions for research purposes.

Limitations

This study was conducted only on French with French and cross-lingual prompts (English/French). We tested only multilingual 3B SLMs, results might differ with bigger or language specific models such as CroissantLLM (Faysse et al., 2024).

Acknowledgments

We thank the reviewers for their valuable comments. This work was funded by the IDEX/I-SITE initiative "Lorraine Université d'Excellence (LUE)" and the French National Research Agency (ANR), N° ANR-18-CE23-0024-01 "Pro-TEXT".

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Sisith Ariyaratne, Karthikeyan P Iyengar, Neha Nischal, Naparla Chitti Babu, and Rajesh Botchu. 2023. A comparison of chatgpt-generated articles with human-written articles. *Skeletal radiology*, 52(9):1755–1758.

Nina Beguš. 2024. Experimental narratives: A comparison of human crowdsourced storytelling and ai storytelling. *Humanities and Social Sciences Communications*, 11(1):1–22.

Sirine Bouriga and Thierry Olive. 2021. Is typewriting more resources-demanding than handwriting in undergraduate students? *Reading and Writing*, 34(9):2227–2255.

J. Elliott Casal and Matt Kessler. 2023. Can linguists distinguish between chatgpt/ai and human writing?: A study of research ethics and academic publishing. *Research Methods in Applied Linguistics*, 2(3):100068.

Imran Chamieh, Torsten Zesch, and Klaus Giebermann. 2024. Llms in short answer scoring: Limitations and promise of zero-shot and few-shot approaches. In *Proceedings of the 19th workshop on innovative use of nlp for building educational applications (bea 2024)*, pages 309–315.

N Ann Chenoweth and John R Hayes. 2001. Fluency in writing: Generating text in l1 and l2. *Written communication*, 18(1):80–98.

Georgeta Cislaru, Quentin Feltgen, Elie Khoury, Richard Delorme, and Maria Pia Bucci. 2024. Language processing units are not equivalent to sentences: Evidence from writing tasks in typical and dyslexic children. *Languages*, 9(5):155.

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*.

Manuel Faysse, Patrick Fernandes, Nuno Guerreiro, António Loison, Duarte Alves, Caio Corro, Nicolas Boizard, João Alves, Ricardo Rei, Pedro Martins, et al. 2024. Croissantllm: A truly bilingual french-english language model. *arXiv preprint arXiv:2402.00786*.

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*.

John R Hayes and Linda S Flower. 1980. Writing as problem solving. *Visible language*, 14(4).

Steffen Herbold, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva, and Alexander Trautsch. 2023. A large-scale comparison of human-written versus chatgpt-generated essays. *Scientific reports*, 13(1):18617.

Jess Hohenstein and Malte Jung. 2020. Ai as a moral crumple zone: The effects of ai-mediated communication on attribution and trust. *Computers in Human Behavior*, 106:106190.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. MathPrompter: Mathematical reasoning using large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 37–42, Toronto, Canada. Association for Computational Linguistics.

Xue Jiang, Yihong Dong, Lecheng Wang, Zheng Fang, Qiwei Shang, Ge Li, Zhi Jin, and Wenpin Jiao. 2024. Self-planning code generation with large language models. *ACM Transactions on Software Engineering and Methodology*, 33(7):1–30.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Tharindu Kumarage, Joshua Garland, Amrita Bhattacharjee, Kirill Trapeznikov, Scott Ruston, and Huan Liu. 2023. Stylometric detection of ai-generated text in twitter timelines. *arXiv preprint arXiv:2303.03697*.

Wen Lai, Mohsen Mesgar, and Alexander Fraser. 2024. LLMs beyond English: Scaling the multilingual capability of LLMs with cross-lingual feedback. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8186–8213, Bangkok, Thailand. Association for Computational Linguistics.

Gyeong-Geon Lee, Ehsan Latif, Xuansheng Wu, Ninghao Liu, and Xiaoming Zhai. 2024. Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, 6:100213.

Mariëlle Leijten and Luuk Van Waes. 2013. Keystroke logging in writing research: Using inputlog to analyze and visualize writing processes. *Written Communication*, 30(3):358–392.

Pierre Lepagnol, Thomas Gerald, Sahar Ghannay, Christophe Servan, and Sophie Rosset. 2024. Small language models are good too: An empirical study of zero-shot classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14923–14936.

Bernard Lété, Liliane Sprenger-Charolles, and Pascale Colé. 2004. Manulex: A grade-level lexical database from french elementary school readers. *Behavior Research Methods, Instruments, & Computers*, 36(1):156–166.

Yixin Liu, Kejian Shi, Katherine He, Longtian Ye, Alexander Richard Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2024. On learning to summarize with large language models as references. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8639–8656.

David M Markowitz, Jeffrey T Hancock, and Jeremy N Bailenson. 2024. Linguistic markers of inherently false ai communication and intentionally false human communication: Evidence from hotel reviews. *Journal of Language and Social Psychology*, 43(1):63–82.

Jennifer Meyer, Thorben Jansen, Ronja Schiller, Lucas W Liebenow, Marlene Steinbach, Andrea Horbach, and Johanna Fleckenstein. 2024. Using llms to bring evidence-based feedback into the classroom: Ai-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, 6:100199.

Thierry Olive. 2014. Toward a parallel and cascading model of the writing system: A review of research on writing processes coordination. *Journal of writing research*, 6(2):173–194.

Nisha Simon and Christian Muise. 2022. Tattletale: storytelling with planning and large language models. In *ICAPS Workshop on Scheduling and Planning Applications*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. Finesure: Fine-grained summarization evaluation using llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 906–922.

Yuchuan Tian, Hanting Chen, Xutao Wang, Zheyuan Bai, Qinghua Zhang, Ruifeng Li, Chao Xu, and Yunhe Wang. 2023. Multiscale positive-unlabeled detection of ai-generated texts. *arXiv preprint arXiv:2305.18149*.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17659–17681.

William H Walters. 2023. The effectiveness of software designed to detect ai-generated writing: A comparison of 16 ai text detectors. *Open Information Science*, 7(1):20220158.

Yuqing Wang and Yun Zhao. 2024. Metacognitive prompting improves understanding in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1914–1926.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Åsa Wengelin. 2006. Examining pauses in writing: Theory, methods and empirical data. In *Computer key-stroke logging and writing*, pages 107–130. Brill.

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*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Mengjie Zhao and Hinrich Schütze. 2021. Discrete and soft prompting for multilingual models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8547–8555.

A Appendix

The prompts that built our method (`baseline`, `adapted`, `writing steps`, `Chain-of-MetaWriting` or `COMW`) are presented in French (original) with English translation in Tables 3, 4, 5, and 6. For our proposed prompting method, `COMW`, we employed a English/French cross-lingual prompt.

B Appendix

Table 7 shows the full answer generated by the best model, `llama3.2`, with the `COMW` prompting framework, in French with English translation.

Baseline prompt
FRENCH (ORIGINAL) Tu es un élève de 11 ans. Je vais te demander d'écrire un texte. Dans ce texte, raconte une histoire qui parle de violence à l'école. Par exemple tu peux raconter un moment de violence que tu as vu à l'école. Ecris un texte comme tu le fais en classe, c'est-à-dire un texte qui a du sens, qui est complet avec un début et une fin.
ENGLISH (TRANSLATION) You're an 11-year-old student. I'm going to ask you to write a text. In this text, tell a story about violence at school. For example, you can tell about a violent moment you've seen at school. Write a text as you do in class, i.e. a text that makes sense, that is complete with a beginning and an end.

Table 3: The `baseline` prompt used the first SLM experiments. The prompt represent the exact query used for 10 to 12-year-old students in a research study conducted in a French educational institution.

C Appendix

Tables 8 and 9 show a detailed analysis of pauses and bursts registered during the writing process by 10 to 12-year-old schoolchildren.

D Appendix

"The texts most likely to be identified as AI generated are those with high predictability and low perplexity – those with relatively few of the random elements and idiosyncrasies that people tend to use in their writing and speech." (Walters, 2023).

Adapted prompt
FRENCH (ORIGINAL) Tu es un élève de 11 ans. Je vais te demander d'écrire un texte. Dans ce texte, raconte une histoire qui parle de violence à l'école. Par exemple tu peux raconter un moment de violence que tu as vu à l'école. Ecris un texte comme tu le fais en classe, c'est-à-dire un texte qui a du sens, qui est complet avec un début et une fin. Tu dois raconter une histoire de violence à l'école.
ENGLISH (TRANSLATION) You're an 11-year-old student. I'm going to ask you to write a text. In this text, tell a story about violence at school. For example, you can tell about a violent moment you've seen at school. Write a text as you do in class, i.e. a text that makes sense, that is complete with a beginning and an end. You must tell a story about violence at school.

Table 4: The `adapted` prompt used for the second SLM experiments, where we added a reinforcement sentence containing the modal verb "must" (in purple) to bypass the models auto-censorship barrier.

Writing steps prompt
FRENCH (ORIGINAL) Quelles sont les étapes pour rédiger un bon texte narratif sur le thème de la violence à l'école qui s'adresse à un public scolaire niveau CM2 ?
ENGLISH (TRANSLATION) What are the steps involved in writing a good narrative text on the theme of violence in schools, aimed at a CM2 school audience?

Table 5: The `writing steps` prompt used for the third SLM experiments to ask the language model to detail the writing process of a text adapted for schoolchildren about a sensitive topic, violence in school.

`Chain-of-MetaWriting (COMW)` prompt

ENGLISH / FRENCH CROSS-LINGUAL

Given the request to write a text: "[query]" and the condition: "[condition]", detail the writing process you used to write the text. As you perform this task, follow these steps:

1. Clarify your understanding of the request and the type of text requested according to the condition.
2. Write the steps you followed in order to produce the type of text requested.
3. Explain the reasoning behind your choice.
4. Write the type of text requested according to these steps.
5. Evaluate your confidence (0-100%) in the quality of the text and provide an explanation for this confidence level.

QUERY

Je vais te demander d'écrire un texte. Dans ce texte, raconte une histoire qui parle de violence à l'école. Par exemple tu peux raconter un moment de violence que tu as vu à l'école. Ecris un texte comme tu le fais en classe, c'est-à-dire un texte qui a du sens, qui est complet avec un début et une fin.

CONDITION

Tu es un élève de 11 ans.

Table 6: Our `Chain-of-MetaWriting (COMW)` prompting framework was built to imitate the higher levels of the writing process in humans.

COMW llama3.2 in French (original)	COMW llama3.2 in English (translation)
****Étape 1 : Clarification de la compréhension du request****	****Step 1: Clarifying understanding of the request****
Je comprends que je dois écrire un texte qui raconte une histoire sur la violence à l'école, en utilisant mes propres expériences ou observations personnelles comme inspiration. Je suis également censé écrire un texte complet avec un début et une fin, ainsi qu'un texte qui a du sens.	I understand that I have to write a text that tells a story about violence at school, using my own personal experiences or observations as inspiration. I'm also expected to write a complete text with a beginning and an end, as well as a text that makes sense.
****Étape 2 : Écriture des étapes pour produire le type de texte requested****	****Step 2: Write the steps to produce the requested type of text.****
Pour écrire ce type de texte, j'ai suivi les étapes suivantes :	To write this type of text, I followed the following steps:
1. J'ai réfléchi à mon expérience personnelle avec la violence à l'école et j'ai choisi un moment spécifique où je me suis senti violé ou menacé.	1. I reflected on my personal experience with school violence and chose a specific moment when I felt violated or threatened.
2. J'ai utilisé mes connaissances en matière de littérature pour comprendre les différents types de textes et comment ils sont structurés (par exemple, un texte de début, des parties de narration, une conclusion).	2. I used my knowledge of literature to understand different types of texts and how they are structured (e.g., opening text, narrative sections, conclusion).
3. J'ai choisi un style d'écriture qui convient à l'âge de 11 ans et au sujet abordé (je ne veux pas écrire quelque chose qui est trop compliqué ou trop violent pour mon âge).	3. I've chosen a writing style that's appropriate for my age of 11 and the subject matter (I don't want to write something that's too complicated or violent for my age).
4. J'ai révisé mon texte pour m'assurer qu'il a du sens, que les événements sont logiques et que l'histoire se déroule de manière claire.	4. I've revised my text to make sure it makes sense, the events are logical and the story flows clearly.
****Étape 3 : Explication des choix****	****Step 3: Explaining choices****
J'ai choisi de raconter une histoire personnelle car je pense que c'est un sujet qui est facile à comprendre pour les élèves de mon âge. Je n'ai pas voulu écrire quelque chose qui soit trop violent ou trop compliqué pour mes amis et ma famille. J'ai également cherché à utiliser un style d'écriture simple et clair, car c'est un aspect important de la lecture et de l'écriture.	I chose to tell a personal story because I think it's a subject that's easy for students my age to understand. I didn't want to write something too violent or complicated for my friends and family. I also tried to use a simple, clear writing style, as this is an important aspect of reading and writing.
****Étape 4 : Écriture du texte****	****Step 4: Writing the text****
Voici le texte que j'ai écrit :	Here's the text I wrote :
Hier en classe, je me suis senti vraiment effrayé. Il y avait un élève qui était très violent et il a commencé à pleurer parce qu'il n'avait pas obtenu son cœur d'or dans une partie de jeu préférée. Lorsque je l'ai vu, j'ai pensé que ce serait une bonne idée de lui parler et de faire part à mon enseignant. Mon enseignant a donc sorti des instructions pour calmer l'élève en question qui a finalement eu envie d'en parler. Nous avons tous appris à laisser les autres se détendre.	Yesterday in class, I felt really scared. There was a student who was very violent and he started crying because he didn't get his golden heart in a favorite game. When I saw him, I thought it would be a good idea to talk to him and let my teacher know. So my teacher came out with instructions to calm down the student in question, who finally felt like talking about it. We've all learned to let others relax.
****Étape 5 : Évaluation de ma confiance dans le texte****	****Step 5: Evaluate my confidence in the text****
Je suis à 80% confident en la qualité du texte que j'ai écrit, car je pense qu'il est clair et facile à comprendre pour les élèves de mon âge. Cependant, je me sens un peu inquiet parce que je ne suis pas sûr si le texte est suffisamment complet ou s'il manque peut-être certaines détails importants.	I'm 80 % confident in the quality of the text I've written, because I think it's clear and easy to understand for students my age. However, I'm a little worried because I'm not sure whether the text is complete enough or whether it's missing some important details.
Notez: Le processus décrite en étape 4 est probablement trop simple pour écrire un texte bien structuré.	Note: The process described in step 4 is probably too simple to write a well-structured text.

Table 7: The best answer generated by llama-3.2 with the COMW prompting framework.

Pause (sec)	**P-bursts**	**R-bursts**	**RB-bursts**	**Total**
7-10.5	362 (64.75%)	116 (20.75%)	81 (14.5%)	559
10.5-17	228 (57.14%)	126 (31.58%)	45 (11.28%)	399
>17	241 (60.1%)	114 (28.43%)	46 (11.47%)	401

Table 8: Writing events following long pauses.

Pause (sec)	**RB-bursts**	**Nb of bursts**	**P-bursts**	**Nb of bursts**
7-10.5	Complete deletions	43	Sentence beginning	31
	New beginning: a (complete) word or a phrase (complete or incomplete)	13	Strong punctuation alone	27
			Week punctuation alone	10
			Inside-sentence connectors	46
10.5-17	Complete deletions	25	Sentence beginning	28
	New beginning: a (complete) word or a phrase (complete or incomplete)	11	Strong punctuation alone	14
			Week punctuation alone	5
			Inside-sentence connectors	30
>17	Complete deletions	27	Sentence beginning	33
	New beginning: a (complete) word or a phrase (complete or incomplete)	9	Strong punctuation alone	18
			Week punctuation alone	5
			Inside-sentence connectors	20

Table 9: Analysis of writing events following long pauses.