A Multi-Modal Interactive Narrative Game for French-Speaking Primary School Learners

Qiao Wang[1], **Adnan Labib**[2], **Thierry Geoffre**[3], **Yixuan Huang**[4], **Zheng Yuan**[5]
[1]Hosei University, Japan
[2]King's College London, United Kingdom
[3]University of Luxembourg, Luxembourg
[4]Waseda University, Japan
[5]University of Sheffield, United Kingdom

Abstract

In this study, we built a Large Language Model (LLM)-powered interactive narrative game for French-speaking primary school learners. The system consists mainly of three modules: an initialization module for story structuring, a generation module for iterative generation of story texts and illustrations based on player choices, and a language scaffolding module for providing contextualized explanations of player-selected textual elements. Text generation is prompted using five curriculum-informed proficiency bands spanning primary to early secondary reading levels. A classroom feasibility study was conducted with 35 students aged 9-10 and two teachers at two primary schools in French-speaking Switzerland. Results showed positive student attitudes towards the stories and illustrations and high perceived ease of use. However, students perceived the game's usefulness for French literacy activities as moderate, with many noting that the story language was below their level. Occasional issues in character consistency and repetitive illustrations across scenes were also mentioned. Teachers highlighted potential pedagogical benefits when the system is combined with additional classroom activities.

1 Introduction

Interactive narrative is a form of game-like storytelling in which narratives unfold dynamically in response to player choices, allowing readers to influence plot progression rather than follow a fixed sequence of events. From *Choose Your Own Adventure*™ books to open-world video games, branching narratives are a central element of interactive fiction that enable readers to change the outcome of a narrative (Walton, 2019; Huang et al., 2024). Such narratives provide increasing engagement and agency by transforming reading into an active, decision-driven process. However, the content of traditional interactive narratives remains limited by creators' perspectives and the scope of pre-authored material (Tian et al., 2024).

The emergence of Large Language Models (LLMs) has renewed interest in interactive narrative by enabling dynamic text generation at scale. Rather than requiring authors to pre-write extensive branching scripts, LLMs can generate story continuations on demand, accommodating player input within narrative constraints defined by the author (Huang et al., 2024; Wang et al., 2025). This shift has substantially reduced authorial overhead and enabled interactive storytelling systems that respond flexibly to user actions. Such LLM-powered interactive narratives also offer pedagogical potential for young readers by creating engaging, adaptive reading experiences. However, existing systems are often English-centric, primarily text-based, not designed for younger users, and lack explicit mechanisms for conditioning linguistic difficulty (e.g., Wu et al., 2024; Xu et al., 2025; Huang et al., 2024; Wang et al., 2026). This creates a gap for systems that can generate age-appropriate reading experiences at different intended difficulty levels.

To address such limitations, we built an LLM-powered multi-modal interactive narrative game for French-speaking primary school learners. The novelty of the system lies in combining: (1) curriculum-informed proficiency conditioning for French story generation, (2) dynamically generated illustrations accompanying story texts, and (3) an in-context language assistant that explains text elements upon player query. The source code is released under the CC BY 4.0 license and is available at `https://github.com/judywq/interactive-narrative-game`. The web application uses proprietary LLMs, whose weights and use are subject to the respective providers' terms of service.

Proceedings of the 1st Workshop on Multi-modality and Multi-lingual Language Learning, pages 65–76
November 9, 2026 ©2026 Association for Computational Linguistics

2 Related Work

Recent years have seen growing interest in the use of LLMs for interactive storytelling, driven by their ability to generate fluent narrative text in response to user input. Existing systems differ in how they structure narrative control, integrate player agency, and manage long-term coherence.

One prominent line of work generates interactive narratives by branching from an existing linear story. WHAT-IF (Huang et al., 2024) introduces a meta-prompting framework that constructs alternative narrative trajectories grounded in a pre-written source text, producing "what-if" timelines that explore counterfactual developments of an original plot. By anchoring generation to a fixed narrative backbone, such systems achieve strong thematic consistency and enable systematic exploration of branching structures, but presuppose the availability of high-quality source stories. Related approaches similarly focus on expanding latent or un-narrated branches within existing story worlds.

A second group of systems emphasizes authorial control through structured narrative representations. Dramamancer (Wang et al., 2026) exemplifies this approach by allowing authors to define story schemas, including characters, events, and constraints, that are then instantiated dynamically during interactive playthroughs using an LLM. Other work explores hybrid approaches that integrate LLM-based generation with concepts from classical AI planning. Though earlier planning-based story generation systems ensured long-term coherence through explicit plot graphs but lacked natural language generation (Meehan, 1977; Young et al., 2013), more recent systems combine LLMs with structured representations such as quest trees or plot graphs to generate interactive fiction worlds, scenes, or branching quests (Ammanabrolu et al., 2019; de Lima et al., 2021; Kumaran et al., 2023; Harmon and Rutman, 2023).

Evaluation of these systems has revealed limitations, particularly with respect to maintaining meaningful agency and narrative coherence over extended interactions (Yi et al., 2025). To address these limitations, researchers have turned to memory-augmented approaches by introducing persistent representations of characters, events, and world states that can be retrieved and updated across interactions, enabling more consistent long-term storytelling (Yi et al., 2025; Nonaka and Perry, 2025). On the other hand, empirical evaluations also reveal systematic tendencies toward positive, low-conflict plot developments and reduced diversity, novelty, and surprise compared to human-authored narratives (Tian et al., 2024; Ismayilzada et al., 2024).

Beyond coherence and engagement, existing LLM-based interactive storytelling systems are limited in their broader applicability. Most systems remain predominantly text-based (Walton, 2019; Huang et al., 2024), offering limited multi-modal support that could aid comprehension. Moreover, prior work has focused largely on English narratives (Huang et al., 2024) and has typically targeted adolescent or adult users (Walton, 2019), implicitly assuming higher reading proficiency, cognitive maturity and digital literacy. For instance, OpenTheater by Xu et al. (2025) requires players to configure detailed story settings and character profiles, and to interact through typed chat to drive the narrative forward. Finally, these systems are generally not designed around pedagogical objectives and lack explicit mechanisms for aligning generated language with player proficiency levels or instructional goals (Wang et al., 2025), which limits their suitability for educational contexts.

In light of these limitations, we set out to develop a multi-modal interactive narrative game for young readers in a non-English linguistic context. We focused on children because interactive storytelling plays an important role in their development, while existing AI-based storytelling and reading technologies still often fall short of parents' expectations (Sun et al., 2024). We selected French as the target language because it is widely taught and is comparatively well supported by LLMs as a high-resource language (Li et al., 2025).

3 System Design and Development

In designing the game content, we prioritized two factors: pedagogical relevance and multi-modal story generation. Pedagogically, we defined the story genre, conditioned generation on intended reading levels, and incorporated language scaffolding. For multi-modal story generation, we incorporated illustrations alongside the story text based on the cognitive theory of multimedia learning (Mayer, 2024; Paivio, 1979), which suggests that combining verbal information (e.g., text) with nonverbal information (e.g., visuals) can support encoding and recall. In addition, given the digital literacy of primary schoolers, we designed player inter-

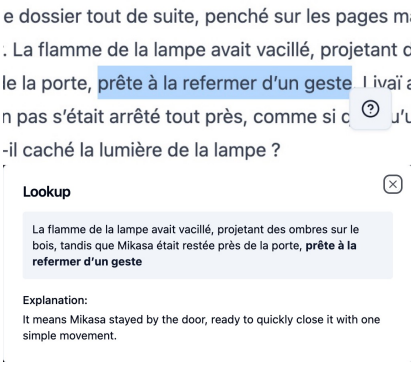


Figure 1: Language query function

action around structured choices rather than free-response chat. For system architecture, we adopted a modular multi-pipeline design (Parnas, 1972) and deployed the system as a web application with a Vue.js frontend and a Django backend.

3.1 Pedagogical Features

Genre selection We focused story generation around the fable genre. Fables are prominent in primary curricula due to their accessible narrative structure, conventional character roles, and explicit moral framing, making them ideal for reading comprehension to young learners. We defined a schema of essential structural elements in fables, which can be found in Appendix A.

Proficiency conditioning We designed five curriculum-informed proficiency bands through consultation with a certified French language teacher. For each band, we curated a dataset of fable texts that were proofread and annotated according to specific curriculum goals. The bands and sample fable texts are used as few-shot examples to condition generation toward the selected reading level (see Appendix B).

Language scaffolding We added a text query feature for players to request explanations of unfamiliar textual elements ranging from single words and phrases to sentences. A player may select the relevant text and click the question mark that subsequently appears. The system then provides a contextualized explanation in French or English (see Figure 1). At the end of a story, queried items are used to generate a recall quiz. The LLM assigns each response a score of 0, 0.5, or 1, and the system calculates a normalized score by dividing the sum of item scores by the number of questions attempted.

3.2 Multi-Modal Story Generation

For story text generation, we selected GPT-5.2[1] as the base LLM, the most advanced model from the GPT series when the study took place. For the structuring of branching narratives, we drew experience from famous commercial interactive narratives such as Detroit: Become Human[2] and Heavy Rain[3]. We base the story generation on a story skeleton, which includes main characters, multiple milestone events and a fixed number of possible endings. The story development to each milestone event progresses in multiple scenes, where players are asked to make decisions towards the end to decide what will happen in the next scene. Each milestone event remains reachable regardless of player choices.

For generating illustrations accompanying each scene, we integrated Gemini-2.5-Flash[4], which excelled in both the quality of images and speed of generation at the time of the study. At each scene, the system passes three types of contextual information to the image generation model: (1) the current scene description, (2) a summary of the story up to the current scene, and (3) the illustration from the previous scene as a reference.

3.3 System Architecture

The system consists of an initialization module, a generation module and a language scaffolding module, as shown in Figure 2. In the initialization module, when a player begins a new game by providing the necessary input, the "Proficiency Benchmarker" pipeline generates five text samples of the same narrative content using the five proficiency-band prompts described in Section 3.1. Figure 5 in Appendix B shows examples of texts at the five bands. The player selects a comfortable level, which is then used as a prompt condition for subsequent text generation. At the same time, the "Story Skeleton Generator" pipeline creates the complete narrative structure (characters, milestones, decision points, and possible endings) based on player input following the fable conventions defined in the LLM prompt.

In the generation module, the "Plot Writer" pipeline iteratively generates narrative scenes based on the story skeleton, selected proficiency level,

[1] https://platform.openai.com/docs/models/gpt-5.2
[2] https://www.quanticdream.com/en/detroit-become-human
[3] https://www.quanticdream.com/en/heavy-rain
[4] https://ai.google.dev/gemini-api/docs/models#gemini-2.5-flash

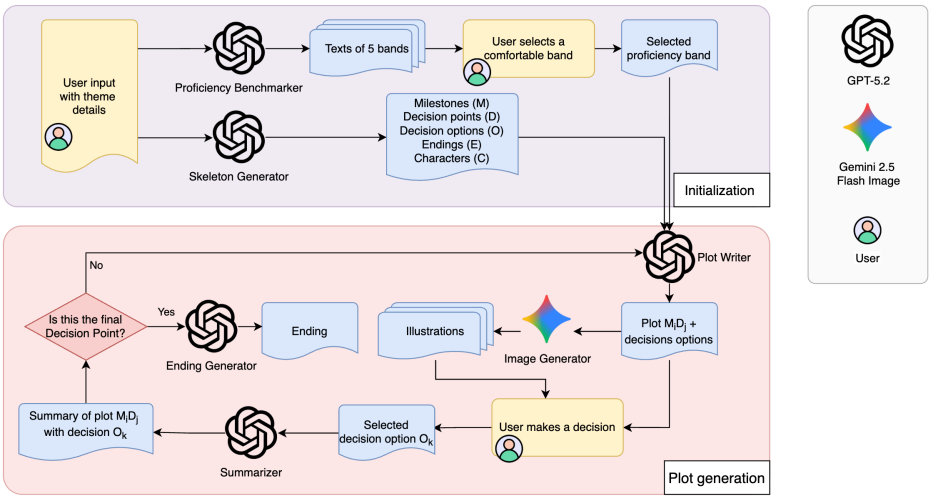


Figure 2: System Architecture

and summary of previous scenes. Subsequently, the "Image Generator" pipeline produces an illustration for each scene using the input mentioned in Section 3.2. After the player makes a decision, the "Story Summarizer" pipeline condenses the scene and choice into a running summary, to be used for next scene generation. When the final decision point is reached, the "Ending Generator" pipeline generates a concluding scene with one of the possible endings.

Independent of story generation, the language-scaffolding module logs each selected textual element and its explanation. If at least one query has been made, the module generates an end-of-story quiz from the queried items and returns the automated normalized score described above.

4 System Validation

We evaluated the system through a user study with 35 primary school students aged 9-10 and two teachers in two intact classes, one at each of two schools in French-speaking Switzerland. The curriculum of the two classes focused on French reading comprehension, which aligned directly with the game's features.

4.1 Participants

The student participants were in their 4th year of primary school, corresponding to level 3 in the five proficiency bands. They were either native French speakers or had been enrolled in French-medium schools since the beginning of their schooling. The study therefore evaluated the system in a French literacy context rather than as a second-language intervention. The teacher participants were experienced teachers with more than 15 years in primary school. Both the teachers and the students are accustomed to utilizing digital tools for learning.

4.2 Experimental Procedures

The validation study lasted for one 50-minute class session in each class. Before the validation study, the researchers provided guidance to the teachers on how to play the game and how to check gameplay logs. At the beginning of each class session, the teacher paired the student participants to reduce the number of simultaneous live-generation requests, distributed shared game accounts, introduced the gameplay mechanics, and asked each pair to test the game by playing a random story. Specifically, the teachers encouraged students to use the customizable "theme" and "scene details" fields to personalize the fable's narrative context, such as "friendship" as the theme and specific locations or characters in the scene details. An example is shown in Figure 3, where a pair requested Luffy, a character from the anime *One Piece*, in the story. Following the initial test, each pair completed one fable game, and the students subsequently completed the player survey. After each session, the corresponding teacher completed the teacher survey based on observations of the class.

4.3 Survey Instruments

The player survey (see Appendix C) was designed based on the Technology Acceptance Model (TAM) (Davis, 1989), focusing on two constructs: Perceived Usefulness (PU) and Perceived Ease of Use



Figure 3: A scene with Luffy from *One Piece*

(PEOU) of the game from the student participants' perspective. Considering the cognitive ability of the students, only three items scaled 1-7 were included under each construct. Open-ended questions were added to allow student participants to freely express their opinions towards the game. The teacher survey (see Appendix D) consisted of five open-ended questions to assess the game's pedagogical integration and its alignment with curriculum standards from the teachers' perspective.

4.4 Data Collection and Analysis

The data collected in the validation study included individual student responses to the player survey, responses from the two teachers, and gameplay logs from the system backend. Because students played in pairs using shared accounts, gameplay logs were analyzed at the pair/account level and cannot be attributed to individual students. For Likert-scale items in the player survey, internal reliability was assessed using Cronbach's alpha for the PU and PEOU scales, followed by descriptive statistics for the individual items. Responses to the open-ended questions in the player and teacher surveys were analyzed using a thematic coding approach by two coders. The gameplay logs were used to complement the survey analysis. For the optional language-scaffolding activity, query records were aggregated by shared account. For each account that completed the follow-up vocabulary quiz, we recorded the number of queries/questions and the automated normalized score produced by the system. Because only six accounts completed the quiz, we report the number of participating accounts, total and per-account query counts, and the mean, median, and range of normalized scores. No inferential tests were conducted for these data.

5 Result

5.1 Player Survey

Likert-scale items Reliability analysis of the two constructs showed good internal consistency for the PU scale (Cronbach's $\alpha = 0.77$) and marginal internal consistency for the three-item PEOU scale (Cronbach's $\alpha = 0.69$; Nunnally and Bernstein, 1994). We therefore interpret PEOU primarily through the individual item statistics. Descriptive statistics shown in Table 1 indicated consistently positive ratings on the PEOU items, with mean scores ranging from 5.66 to 5.80, suggesting that students generally perceived the game as easy to learn, clear to interact with, and easy to understand.

In contrast, the perceived usefulness (PU) items received more moderate ratings, with mean scores ranging from 4.09 to 4.60 and greater variability across participants, indicating more diverse perceptions regarding the learning value of the game.

Open-ended questions Thematic analysis of student comments revealed three important themes. First, regarding the stories, most students expressed positive attitudes (N=24), while several students expressed negative views (N=3), and a small number mentioned technical issues (N=2). One respondent comment stated "only the beginning of a story", indicating network issues in generating scenes after the story was initialized. This comment was consistent with the two observed game freezes during live generation.

Second, regarding the illustrations, the students generally liked the illustrations (N=23), with a small number of neutral responses (N=5). Individual expressions included "beautiful", "pretty", "cool", and "too good". Figure 6 in Appendix E shows a student's drawing of their favorite illustration in the player survey. Meanwhile, three students pointed out occasional inconsistent characters in illustrations (e.g., "one time I asked for a hare it gave me a goat"). Another student also commented that the illustrations looked too similar between scenes. An example is shown in Figure 4. The reason is that most of the story events took place in the same location and thus only minor elements changed between scenes.

Third, regarding language difficulty, 14 students described the game as "easy" or "too easy", while 9 reported fair difficulty. An examination of gameplay logs revealed that the majority of gameplay occurred at level 2, although the participants' expected curriculum band was level 3. This pattern indicates a preference for a lower level of challenge in the self-selection interface.

Construct	Item	Statement	Mean	SD
PU	PU1	Using the game in my French class may enable me to learn to better read and understand texts.	4.60	1.75
PU	PU2	Using the game may improve my French performance.	4.09	1.79
PU	PU3	I find the game useful in learning French.	4.11	2.08
PEOU	PEOU1	Learning to play the game was easy for me.	5.77	1.57
PEOU	PEOU2	My interaction with the game was clear and understandable.	5.66	1.53
PEOU	PEOU3	I found the game clear and understandable.	5.80	1.64

Table 1: Descriptive statistics for the TAM questionnaire items (N = 35).



Figure 4: Similar illustrations between scenes

Pair/account ID	No. of queries/ questions	Automated normalized score
1	4	0.875
2	4	0.875
6	3	0.67
7	1	0.50
9	2	0.50
24	3	0.67

Table 2: Language-query use and automated normalized vocabulary quiz scores for shared accounts that completed the quiz ($n = 6$).

5.2 Vocabulary Quiz Results

Six shared accounts used the language-query function to look up unfamiliar words or phrases and completed the corresponding vocabulary quiz. As shown in Table 2, these accounts made 17 queries in total, with between one and four queries per account. Their automated normalized quiz scores ranged from 0.50 to 0.875 ($M = 0.682$, median $= 0.67$).

5.3 Teacher Survey

The thematic analysis of the teacher survey indicates that the stories are coherent and pedagogically beneficial. The teachers confirmed that the stories maintain strong logical consistency, characterized by clear "cause-consequence" chains. For example, a character's failure is directly linked to specific choices, such as refusing help or acting out of pride. Pedagogically, a notable result is the system's role as an instructional "trigger." One teacher identified the choice-based format as ideal for practicing "anticipation" (predicting outcomes) and "textual justification" (finding evidence for character feelings). Both teachers mentioned that the game text can also be viewed as a "base draft" for writing workshops, where students can move from the digital environment to paper-based activities, such as writing alternative endings or transforming narrative into dialogue. While one teacher described the system's literary style as "flat" and "repetitive", this was framed as a pedagogical benefit as well. For "struggling" readers, the predictable syntax acts as a supportive scaffold. For advanced students, "text improvement" activities may be implemented where pupils may enhance the game's "flat" style with more vivid descriptions and humor.

6 Discussion

Overall, the validation study suggests that the game was well received by the students and teachers, particularly due to its engaging stories and illustrations. One limitation identified by students was the occasional inconsistency and repetitiveness of the generated illustrations. At the time this study was conducted, maintaining character consistency across sequential images remained a well-recognized limitation of visual models (Avrahami et al., 2024; Shen et al., 2025). Although reference images were used to improve character consistency, inconsistencies still occasionally occurred. Moreover, the system's reliance on reference images from previous scenes introduced another trade-off, namely, repetitive illustrations across scenes. Future iterations may introduce location transitions at the skeleton generation stage, for example, by adding a class for locations and marking when the narrative moves

from one place to another or when the environment should change. In this way, the corresponding images can be generated with richer variation.

The classroom deployment also exposed a trade-off between live generation and operational reliability. Although live calls enabled personalization, simultaneous text and image generation exceeded the available school network capacity and caused two games to freeze even after students were paired. A hybrid deployment could pre-generate and teacher-vet common story skeletons, scenes, and illustrations across the five bands, cache frequently used assets, and reserve live generation for constrained personalization and contextual explanations. Such a design would reduce latency and exposure to unreviewed visual inconsistencies while preserving parts of the interactive experience. A local fallback to a cached scene would also allow play to continue when a generation request fails.

Pedagogically, students perceived the game as only moderately useful for French literacy activities. One possible reason is that French was already the students' native language or language of instruction, and many selected proficiency bands below their expected curriculum level. Future iterations could use a brief baseline reading task or a teacher recommendation to establish the starting band, while still allowing students to adjust the level. The system could then recommend text slightly above the established baseline, consistent with the principle of comprehensible challenge expressed in Krashen's $i + 1$ Hypothesis (Krashen, 1977; Loschky, 1994). Whether this design would benefit second-language learners remains an open question requiring evaluation with an L2 population.

The vocabulary quiz logs provide an exploratory account of how the language-scaffolding feature was used during play. Across the six shared accounts that completed the activity, no automated normalized score was below 0.50, and two scores reached 0.875. The scores also followed a descriptive gradient: accounts with four queries scored 0.875, those with three queries scored 0.67, and those with one or two queries scored 0.50. This pattern may reflect differences in engagement, repeated exposure, prior proficiency, or collaboration within pairs. Because the logs are account-level and the LLM grader was not validated against human judgments, the pattern should not be interpreted as evidence of individual vocabulary learning.

The low uptake of the query function is itself informative for system design. Fourteen students described the game as "easy" or "too easy", and gameplay logs showed that many selected a proficiency level below their expected level. Students reading text that contained few unfamiliar words would have little reason to initiate a vocabulary query. Uptake may also have been reduced by the optional, user-initiated design: students may not have noticed the feature, may have preferred to preserve the flow of the story, or may have regarded a separate quiz as additional effort. Future versions could recommend an initial level using a brief proficiency check, adapt difficulty when an account records few queries, and make word support more discoverable through unobtrusive in-text prompts. A one-click definition followed by a short retrieval prompt could integrate practice into the narrative instead of requiring a separate activity. Reintroducing queried words in later scenes and in a delayed quiz would provide multiple opportunities to encounter and recall the words. Finally, finer-grained logs should record whether students saw, opened, and completed each stage of the feature, allowing future studies to distinguish lack of vocabulary need from problems of discoverability, interruption, or quiz completion.

Regarding generalizability, although the current version focuses on a specific genre with French-speaking primary school students, the system provides a flexible, language-agnostic architecture for interactive narratives that may be extended beyond the present use case. The underlying narrative architecture supports alternative narrative schemas, for example, for detective or science-fiction stories, by adjusting the elements that define each genre. Because the architecture is not tied to a particular language, narrative generation and proficiency-conditioning prompts can be adapted for other linguistic contexts. The growing availability of multilingual large language models further facilitates deployment across diverse language populations. The "Proficiency Benchmarker" may also be adapted to other proficiency frameworks, such as the Common European Framework of Reference for Languages (Council of Europe, 2001); however, the resulting outputs would require independent validation in each new context.

7 Conclusion

In this study, we designed, implemented, and conducted a classroom feasibility evaluation of a

multi-modal, LLM-powered interactive narrative game for French-speaking primary school learners. Findings from the classroom study indicate high perceived ease of use, positive engagement with the game system, and potential pedagogical value when the game is combined with other classroom activities. Areas for refinement in reading-level calibration, technical reliability, language scaffolding, and visual diversity were also identified. The system provides a flexible architecture for LLM-powered interactive narratives with curriculum-informed support for different intended reading levels.

Limitations and Ethical Considerations

This study has several limitations that should inform the interpretation of our results. First, data were collected from two intact classes, one at each of two schools, during one 50-minute session per class and with a limited sample size, limiting the generalisability of findings across populations, instructional contexts, and institutional settings. The participants were native French speakers or had received French-medium schooling since Grade 1; consequently, the findings concern French literacy activities in this context and cannot establish usefulness for second-language learners. Future work should replicate the study across additional schools, sessions, and learner populations.

Second, the study evaluated short-term feasibility, usability, and participant perceptions rather than learning, transfer, or retention. It included no pretest, comparison condition, or delayed assessment. In addition, the three-item PEOU scale showed marginal internal consistency ($\alpha = 0.69$), so the ease-of-use findings should be interpreted primarily at the item level rather than as a robust latent construct.

Third, the five curriculum-informed bands were implemented through teacher consultation, annotated examples, and prompting, but generated scenes were not independently audited for lexical, syntactic, or curriculum-band compliance. Nor did the study systematically evaluate the grammaticality, age appropriateness, or cross-scene coherence of generated French text. The bands should therefore be understood as generation targets rather than validated proficiency classifications.

Fourth, the vocabulary-quiz evidence is exploratory. Students played in pairs using shared accounts, so query behavior and quiz scores cannot be attributed to individual students. Only six accounts used the language-query function and completed the optional quiz, producing a small, self-selected subgroup. Quiz length and content varied with the words queried, and the automated 0/0.5/1 grades were not validated against human ratings. No criterion-referenced mastery threshold was specified. The normalized scores therefore cannot establish vocabulary mastery, learning gains, retention, or a causal relationship between query frequency and performance.

Finally, the live-generation design created technical and safety limitations. Simultaneous text and image requests exceeded the available school network capacity and caused two games to freeze. The study also did not systematically evaluate input or output moderation, model bias, or harmful and age-inappropriate content, even though the theme and scene-details fields accepted free text. Teacher supervision reduced immediate classroom risk but does not substitute for technical safeguards. Future deployments should combine age-appropriate input filtering, output moderation, restricted topic options, teacher review controls, audit logging, and cached fallbacks for failed generation requests.

From an ethical standpoint, working with children necessitates strong protections for privacy and autonomy. We adhered to GDPR principles (European Parliament and Council of the European Union, 2016), minimised collection of identifiable information, and secured research data. Informed consent and age-appropriate explanations ensured that participation was voluntary. Teachers were present during all activities, and learners were informed when content was AI-generated to support transparency and agency.

Generative models also exhibit non-deterministic outputs and sensitivity to prompt variation and provider updates. We documented prompts and model configurations to support reproducibility, but exact regeneration of outputs cannot be guaranteed.

Acknowledgments

The authors would like to thank the students and teachers who participated in the user study. Special thanks also go to Mr. Lin Shijiang for providing the computing resources necessary for this research.

References

Prithviraj Ammanabrolu, Ethan Tien, Wesley Cheung, Zhaochen Luo, William Ma, Lara J. Martin, and Mark O. Riedl. 2019. Guided neural language generation for automated storytelling. In *Proceedings of the Second Workshop on Storytelling*, pages 46–55, Florence, Italy. Association for Computational Linguistics.

Omri Avrahami, Amir Hertz, Yael Vinker, Moab Arar, Shlomi Fruchter, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. 2024. The chosen one: Consistent characters in text-to-image diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12. ACM.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, Cambridge, UK.

Fred D. Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3):319–340.

Edirlei Soares de Lima, Bruno Feijó, and Antonio L. Furtado. 2021. Adaptive branching quests based on automated planning and story arcs. In *Proceedings of the 20th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)*, pages 9–18.

European Parliament and Council of the European Union. 2016. Regulation (EU) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union.

Sarah Harmon and Sophia Rutman. 2023. Prompt engineering for narrative choice generation. In *Interactive Storytelling*, volume 14383 of *Lecture Notes in Computer Science*, pages 208–225. Springer Nature Switzerland.

Runsheng Huang, Lara J. Martin, and Chris Callison-Burch. 2024. WHAT-IF: Exploring branching narratives by meta-prompting large language models. *Preprint*, arXiv:2412.10582.

Mete Ismayilzada, Claire Stevenson, and Lonneke van der Plas. 2024. Evaluating creative short story generation in humans and large language models. *Preprint*, arXiv:2411.02316.

Stephen D. Krashen. 1977. The input hypothesis. In *On TESOL '77: Teaching and Learning English as a Second Language: Trends in Research and Practice*.

Vikram Kumaran, Jonathan Rowe, Bradford Mott, and James Lester. 2023. SceneCraft: Automating interactive narrative scene generation in digital games with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 19, pages 86–96. Association for the Advancement of Artificial Intelligence.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2025. Language ranker: a metric for quantifying llm performance across high and low-resource languages. AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Lester Loschky. 1994. Comprehensible input and second language acquisition. *Studies in Second Language Acquisition*, 16(3):303–323.

Richard E. Mayer. 2024. The past, present, and future of the cognitive theory of multimedia learning. *Educational Psychology Review*, 36(1):8.

James R. Meehan. 1977. TALE-SPIN, an interactive program that writes stories. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, volume 1, pages 91–98.

Hiroshi Nonaka and K. E. Perry. 2025. Evaluating LLM story generation through large-scale network analysis of social structures. *Preprint*, arXiv:2510.18932.

Jum C. Nunnally and Ira H. Bernstein. 1994. *Psychometric Theory*, 3rd edition. McGraw-Hill, New York.

Allan Paivio. 1979. *Imagery and Verbal Processes*. Psychology Press, New York.

David L. Parnas. 1972. On the criteria to be used in decomposing systems into modules. *Communications of the ACM*, 15(12):1053–1058.

Fei Shen, Hu Ye, Sibo Liu, Jun Zhang, Cong Wang, Xiao Han, and Yang Wei. 2025. Boosting consistency in story visualization with rich-contextual conditional diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6785–6794. Association for the Advancement of Artificial Intelligence.

Yuling Sun, Jiaju Chen, Bingsheng Yao, Jiali Liu, Dakuo Wang, Xiaojuan Ma, Yuxuan Lu, Ying Xu, and Liang He. 2024. Exploring parent's needs for children-centered ai to support preschoolers' interactive storytelling and reading activities. 8(CSCW2).

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, Miami, Florida, USA. Association for Computational Linguistics.

Nick Walton. 2019. AI Dungeon. Developed by Latitude.

Qiao Wang, Adnan Labib, Robert Swier, Michael Hofmeyr, and Zheng Yuan. 2025. GenQuest: An LLM-based text adventure game for language learners. *Preprint*, arXiv:2510.04498.

Tiffany Wang, Yuqian Sun, Yi Wang, Melissa Roemmele, John Joon Young Chung, and Max Kreminski. 2026. Design techniques for LLM-powered interactive storytelling: A case study of the Dramamancer system. *Preprint*, arXiv:2601.18785.

Weiqi Wu, Hongqiu Wu, Lai Jiang, Xingyuan Liu, Jiale Hong, Hai Zhao, and Min Zhang. 2024. From role-play to drama-interaction: An LLM solution. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3271–3290. Association for Computational Linguistics.

Tianyang Xu, Hongqiu Wu, Weiqi Wu, and Hai Zhao. 2025. OPEN-THEATRE: An open-source toolkit for LLM-based interactive drama. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 453–460. Association for Computational Linguistics.

Qiang Yi, Yangfan He, Jianhui Wang, Xinyuan Song, ShiYao Qian, Xinhang Yuan, Yi Xin, Yijin Wang, Jingqun Tang, Yuchen Li, Junjiang Lin, Hongyang He, Zhen Tian, Tianxiang Xu, Keqin Li, Kuan Lu, Menghao Huo, Jiaqi Chen, Miao Zhang, and 2 others. 2025. SCORE: Story coherence and retrieval enhancement for AI narratives. *Preprint*, arXiv:2503.23512.

R. Michael Young, Stephen G. Ware, Bradly A. Cassell, and Justus Robertson. 2013. Plans and planning in narrative generation: A review of plan-based approaches to the generation of story, discourse and interactivity in narratives. *Sprache und Datenverarbeitung, Special Issue on Formal and Computational Models of Narrative*, 37(1-2):41–64.

A Story elements designed for fables

```
# Story Elements Reference

The story should incorporate the following classic
narrative elements:

1. Stable beginning
e.g., Known environment (maison, forêt, village)

2. Desire / problem
e.g., Wanting something
(go to the ball, eat, explore, help)

3. Obstacle
e.g., Villain, rule, fatigue, danger, limitation

4. Repeated attempts
e.g., Often 2 or more times with increasing intensity

5. Resolution
e.g., Help, cleverness, cooperation

6. Explicit moral
e.g., Stated directly, not inferred

# Character Design Patterns

- Protagonist: Childlike, curious, imperfect
- Helper: Animal, fairy, friend, object
- Obstacle: Not morally complex
- Collective (optional): Group solves problem together
```

B Five proficiency bands

```
# Level Guidelines

Level 1
Description:
Very short sentences. One concrete action per sentence.
Only high-frequency, concrete vocabulary.
Simple connectors (et, puis, mais, alors).

Example:
Le petit garçon entre dans la forêt. Il voit une
maison. Il a faim alors il frappe à la porte.

Level 2
Description:
Simple narration with basic subordination
(quand, parce que, comme).
Limited description and explicit cause-effect.

Example:
Quand le petit garçon entre dans la forêt, il voit
une petite maison. Comme il a faim, il frappe à
la porte.

Level 3
Description:
More compact language with participial
structures and varied vocabulary.
Do not introduce new events.

Example:
En entrant dans la forêt, le petit garçon aperçoit
une petite maison isolée. Après un moment,
il décide d'entrer.

Level 4
Description:
More fluent narration with light abstract nouns
(e.g., silence, hésitation), while keeping the
same events.

Example:
En entrant dans la forêt, le petit garçon aperçoit
une maison isolée. Après un moment, il décide
d'entrer.

Level 5
Description:
Condensed literary language with richer
vocabulary. Use light metaphor without adding
new narrative content.

Example:
Au cœur de la forêt, le petit garçon distingue
une maison solitaire. La faim, plus forte que la
prudence, l'incite à frapper.
```

C Player Survey (administered in French)

Instructions: Please read each statement carefully and indicate how much you agree or disagree.

Response scale:

7-point Likert scale:

1 = Strongly disagree;
2 = Disagree;
3 = Slightly disagree;
4 = Neither agree nor disagree;
5 = Slightly agree;
6 = Agree;
7 = Strongly agree.

Perceived Usefulness (PU)

1. Using the game in my French class may enable me to learn to better read and understand texts.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7

2. Using the game may improve my French performance.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7

3. I find the game useful in learning French.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7

Perceived Ease of Use (PEOU)

4. Learning to play the game was easy for me.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7

5. My interaction with the game was clear and understandable.

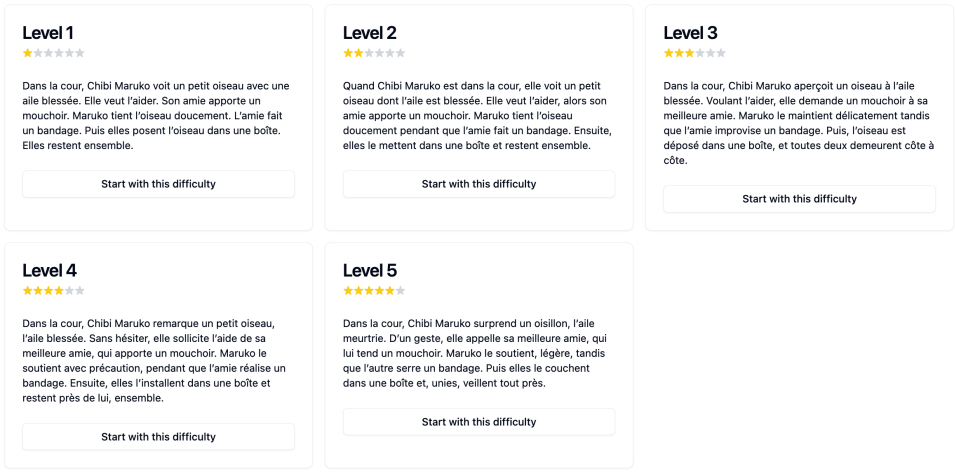


Figure 5: Text samples of the five proficiency bands

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7

6. I found the game clear and understandable.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7

Open-Ended Questions

7. Draw what you liked about the game in the space below.

Drawing space

8. Did you like the story of the game?

_____

9. What do you think about the illustrations in the game?

_____

10. Were the stories too difficult, too easy, or just right for you?

_____

11. Would you recommend this game to your classmates or to your friends at school?

_____

12. Do you have any other suggestions or comments about the game?

_____

D Teacher Survey

1. **Pedagogical Coherence:** Do you think that the generated narratives maintain sufficient logical flow for regular classroom use? How well did the levels of difficulty adapt to your students' needs?
2. **Written Production:** Can you imagine a way to use the game to help mastery of written expression and textual genres?
3. **Social-Emotional Learning:** Could the game be used to promote the development of empathy and social competencies through narrative engagement? If so, in what ways?
4. **Oral Communication:** Does the game provide opportunities for verbal interaction and storytelling?
5. **Creativity:** To what degree does the game foster original thought and narrative divergent thinking?

E Example of students' favorite illustrations



Figure 6: A student's favorite illustration