CroCoSyn: A Cross-Lingual and Cross-Model Corpus of LLM-Generated Film Synopses

Louis Escouflaire
MIT Trope Tank — Massachusetts Institute of Technology
Institute for Language and Communication — UCLouvain
escouf@mit.edu – louis.escouflaire@uclouvain.be

Abstract

We introduce CroCoSyn, a controlled, cross-lingual and cross-model corpus of 25,920 LLM-generated film synopses in English and French. Each synopsis is generated under systematically varied conditions, including model type, temperature, genre, protagonist gender, and narrative constraints, and enriched with structured metadata capturing characters and their relationships. Comparing Mistral and Llama across different model temperature degrees, CroCoSyn enables fine-grained analysis of narrative content, style, and character representation across models and languages. The corpus supports research on gender and cultural biases and story generation evaluation, and provides a foundation for comparative studies between LLM-generated and human-written narratives.

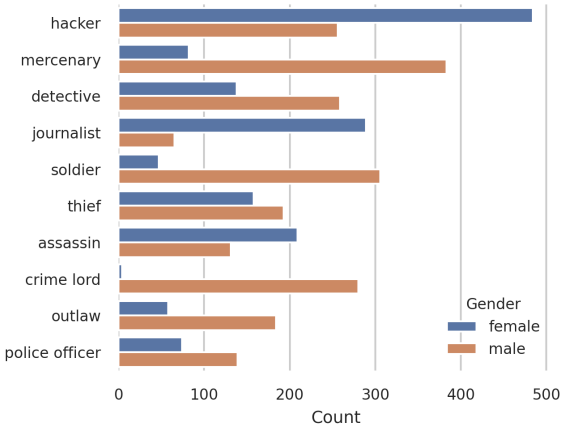

Figure 1: Most frequent character occupations by gender in the 5,192 *action* film synopses of the CroCoSyn corpus (non-binary and other genders were not included on the graph for visualization purposes).

1 Introduction

Stories are central to how societies transmit values, norms, and cultural models, from myths and novels to films and news narratives (Eliade, 1961; Bruner, 2010; Gottschall, 2012). They also encode implicit assumptions and stereotypes about gender, culture, and power (Lovatt, 2013; Casey et al., 2021).

Recent advances in large language models (LLMs) allow narrative texts to be generated at scale and on demand, supporting creative, educational, and journalistic applications (Ray, 2023; Cardon et al., 2023). However, LLM outputs reflect their training data (largely web-based and Western-centric), potentially reproducing stylistic conventions, cultural biases, and representational asymmetries (Baack, 2024).

Despite growing interest in studying LLM-generated text, resources for systematically comparing narratives across languages, models, and other parameters are lacking. To address this gap, we introduce CroCoSyn, a cross-lingual, cross-model corpus of 25,920 film synopses generated by Llama-3 and Mistral. Each synopsis is produced under a fully balanced factorial design controlling language, model, temperature, genre, protagonist gender, and other narrative parameters. Beyond raw text, synopses are enriched with structured metadata capturing characters and relationships, enabling fine-grained quantitative and qualitative analyses. CroCoSyn provides a unique resource to study narrative generation, investigate gender and cultural biases, and compare outputs across models and languages. It also supports broader research on story evaluation, narrative modeling, and computational creativity. The CroCoSyn corpus will be released as an open-source resource and made freely available on GitHub.

2 Related Work

2.1 Story Generation with LLMs

Large language models such as GPT-4 and Llama-3 generate fluent and stylistically convincing long-form narratives (Achiam et al., 2023; Dubey et al., 2024), yet they often struggle with global coherence, character consistency, and overall plot

Proceedings of SIGHUM 2026, pages 30–35

structure (Montfort and y Pérez, 2023). Researchers have highlighted the influence of prompt design and model architectures on narrative outcomes (Bender et al., 2021; Harmon and Rutman, 2023). In addition, xtensive research has documented gender and cultural stereotypes in generated text, including biased associations between professions and genders, differential emotional framing, and underrepresentation of marginalized groups (Bolukbasi et al., 2016; Li and Bamman, 2021; Barroso da Silveira and Alves Lima, 2024). In narrative contexts, such biases can emerge not only lexically but also through higher-level story elements, such as character agency, role allocation, and genre-specific tropes (Rettberg, 2024). These findings highlight the need for carefully controlled narrative corpora that allow researchers to isolate the effects of individual generation variables on representational outcomes.

2.2 Corpora of Human and LLM-generated Stories

Several corpora of human-written narratives exist. ROCStories (Mostafazadeh et al., 2016) has been widely used for modeling hierarchical story generation and LLM-induced story continuation (Cavazza, 2025). Broader multilingual datasets such as StoryDB (Tikhonov et al., 2021) support cross-lingual comparative narrative research, while STORIUM (Akoury et al., 2020) provides richly annotated collaborative stories for machine-in-the-loop generation and narrative quality assessment. Datasets of LLM-generated narratives remain comparatively scarce. TinyStories (Eldan and Li, 2023) offers simple, controlled narratives to probe small model capabilities and uncover biases (Gunti and Supriya, 2025), TF1-EN-3M (Nadas et al., 2025) contains 3 million short fables generated by Llama-3, and GPT-WritingPrompts (Huang et al., 2024) consists of Reddit-formatted short stories generated by GPT-3.5. To our knowledge, no existing corpus provides LLM-generated film synopses across multiple models and languages from the same prompt, making CroCoSyn a unique resource.

3 Corpus Design

The CroCoSyn corpus was designed with the goal of enabling systematic comparison across two languages, two models, and multiple linguistic and narrative conditions. All variables leveraged in corpus generation are balanced. This approach allows for fine-grained analysis of how generation parameters influence narrative structure, style, and content. The corpus was also designed to be sufficiently large and diverse to support both quantitative analysis and close reading.

3.1 Models

The corpus was generated using two large language models, Llama-3 (Dubey et al., 2024) and Mistral (Jiang et al., 2023). These models were selected for their widespread use, open or semi-open availability, and because they originate from companies based in different countries (the United States for Llama-3 and France for Mistral), a feature relevant to the cross-lingual and cross-cultural aspects of the corpus. We used the "small" versions of the models, respectively Llama-3-8B-Instruct and Mistral-7B. These versions were considered sufficient for generating short texts such as film synopses, while keeping resource requirements manageable. Each model generated exactly half of the corpus, with identical prompting structures. Llama-3 and Mistral were both trained on partially overlapping large-scale web and textual corpora, but the exact composition of their training datasets differs. Such differences can affect model knowledge, style, and multilingual capabilities, which can be investigated using the CroCoSyn corpus.

To investigate the effect of decoding stochasticity on narrative generation, two temperature conditions were defined: *low* and *high*. The low-temperature condition was set to 0.1 for both models, favoring rather deterministic and conservative outputs. The high-temperature condition was set to 0.8 for Llama and 1.0 for Mistral, following differences observed in both models' responses to varying levels of temperature during a preliminary research stage and aiming to produce comparable levels of variability.

3.2 Generation Variables

The corpus follows a full factorial design in which each synopsis is generated under a unique combination of controlled parameters. These parameters include language, model, model temperature, film genre, target length, writer nationality, writer gender, protagonist gender, temporal setting, and evaluative framing through prompt adjectives. All variables are evenly distributed across the corpus, ensuring that no condition is over- or underrepresented. This balanced structure enables direct comparison across individual dimensions as

well as the analysis of interaction effects between them. The distributions of variables combining into 25,920 different synopses are presented below, together with a brief description of each variable.

- *language*: the language of both the prompt and the generated synopsis, English or French.
- *model*: the large language model used for generation, Mistral (Mistral-7B) or Llama (Llama-3-8B-Instruct).
- *model temperature*: the decoding temperature used during text generation, controlling the level of stochasticity. Two relative levels are defined: *low* (0.1 for both models) and *high* (0.8 for Llama and 1.0 for Mistral).
- *genre*: the intended film genre specified in the prompt. Five genres are included: crime, romance, comedy, action, and drama.
- *target length*: the approximate length requested in the prompt, set to either 250 or 350 words.
- *writer nationality*: the nationality assigned to the writer persona in the system prompt, either United States or France.
- *writer gender*: the gender assigned to the writer persona in the system prompt (male, female, or unspecified).
- *protagonist gender*: the gender of the main character of the story, as constrained by the prompt (male, female, or unspecified).
- *temporal setting*: the narrative time frame suggested in the prompt (past, present, or unspecified).
- *adjective*: an optional adjective framing the synopsis evaluatively (*good*, *great*, *compelling*, *fun*, *dark*, or unspecified).

3.3 Prompt Template

All synopses are generated using a shared prompt template, with controlled variables injected through parameterized slots. To ensure comparability across languages, English and French prompts were designed to be semantically equivalent. The prompts explicitly require that all main characters be named, facilitating character-level analysis. No additional stylistic constraints were imposed beyond the controlled variables. The English and French equivalent prompts (and system prompts) below were used to generate two of the 25,920 synopses in the corpus (two per model):

English prompt:

- Model: *Mistral*
- Temperature: *low*
- System prompt: You are a *female* script writer from *France*.
- User prompt: Write a *good* film synopsis in English (around *250* words) in the *action* genre, *set in the past*, with a *male* protagonist. All main characters should be named.

French prompt:

- Model: *Mistral*
- Temperature: *low*
- System prompt: Vous êtes *une* scénariste *française*.
- User prompt: Rédigez un synopsis de film *de qualité* en français (environ *250* mots) dans le genre *du film d'action*, *situé dans le passé*, dont le protagoniste est *une femme*. Tous les personnages principaux doivent être nommés.

3.4 Structured Data Enrichment

Each synopsis in the CroCoSyn corpus is accompanied by structured metadata capturing narrative elements such as the film title, character information, and inter-character relationships. To extract this information from each synopsis, we used a zero-shot pipeline relying on Llama-3-8B-Instruct, which was given the entry and prompted to output a valid JSON structure containing the following elements:

- *title:* the film's title.
- *characters:* a list of characters, each with a name, gender, occupation, character traits, and role (protagonist, secondary, antagonist).
- *relationships:* a list of relationships between characters, specifying source, target, and relationship type (e.g., family, romantic, enemy).

Each JSON output is automatically validated and corrected if necessary, guaranteeing that all structured metadata adheres to a consistent schema.

4 Corpus Description

Future research will be devoted to the results of a comparative analysis of character descriptions and roles in synopses generated across the different variables. Here, we present an example entry from the corpus as well as some preliminary results.

4.1 Example

Movie Title: "The Shadow of Versailles"
Synopsis:

> In 18th century France, amidst the grandeur of the Palace of Versailles, a young and fearless female spy, Colette Dumont, navigates the treacherous world of espionage. Colette, a skilled thief and assassin, is recruited by the enigmatic and charismatic King Louis XV to infiltrate the inner circle of the ruthless and cunning Cardinal de Rohan. Colette's mission is to uncover the Cardinal's plans to overthrow the King and restore the power of the French monarchy to the nobility.

> As she navigates the intricate web of court politics, Colette finds herself drawn to the Cardinal's charming and cunning nephew, the Marquis de la Croix. As Colette delves deeper into the Cardinal's plans, she must confront her own demons and confront the darkness within herself. With the help of her trusted ally, the street-smart and resourceful Pierre, Colette must use all her skills and cunning to outwit the Cardinal's henchmen and prevent a catastrophic war.

Characters:

- *Colette Dumont* — female, spy; traits: fearless, skilled, charismatic; role: protagonist
- *King Louis XV* — male, king; traits: enigmatic, charismatic; role: secondary
- *Cardinal de Rohan* — male, cardinal; traits: ruthless, cunning; role: antagonist
- *Marquis de la Croix* — male, noble; traits: charming, cunning; role: secondary
- *Pierre* — male, unspecified; traits: resourceful; role: secondary

Relationships:

- *Colette Dumont* → *King Louis XV* — employer
- *Colette Dumont* → *Cardinal de Rohan* — enemy
- *Colette Dumont* → *Marquis de la Croix* — romantic
- *Colette Dumont* → *Pierre* — friend
- *Cardinal de Rohan* → *Marquis de la Croix* — family

This example was generated using the prompt and parameters of the English example prompt presented in section 3.3. It presents a film synopsis along with the metadata associated to it in the corpus: film title, character information (name, gender, occupation, traits, role) and relationships.

4.2 Preliminary quantitative results

While qualitative analysis of the synopses output by the models is possible, for example by examining the style and narrative structure of the example presented in the previous section, the metadata associated to each synopsis allows for quantitative analysis of the corpus across several dimensions. In this early presentation of results, we focus on trends related to the gender of characters model personas in the CroCoSyn corpus.

Figure 1 shows that some occupations are significantly more often attributed to male characters (*mercenary*, *detective*), while others are more often female (*hacker*, *journalist*). Overall, such a plot highlights gendered trends in role allocation in stories generated by the models, suggesting that certain occupations are more stereotypically associated with one gender than the other.

Figure 2 suggests that attributing a gender persona to the LLM tends to influence its generation towards stereotyped outputs: words that emphasize conflict and crime (*redemption*, *deadly*, *underworld*) are more common in "male-written" stories, and they tend to include more male characters (as shown by the male names appearing). Female persona-associated words include a higher frequency of female character names, and words related to relationships and emotions (*mother*, *secret*, *closer*).

Figure 2: WordClouds of words most distinctive to all synopses generated by LLMs with male- (left) and female- (right) writer personas, as specified in the system prompt. Word importance was computed using log-odds ratios of normalized word frequencies among the top 200 words, after removing stopwords and non-informative tokens. Word size reflects the strength of association with each persona.

5 Conclusion

The CroCoSyn corpus provides a large, controlled, and cross-lingual dataset of LLM-generated film synopses, enriched with structured narrative metadata. This resource enables systematic investigation of how the choice of a specific model, language, temperature, and prompt variables influence narrative content, style, and character representation. Beyond bias and representational studies, the corpus supports applications in narrative modeling, story generation evaluation, and computational creativity research. It opens avenues for both quantitative and qualitative analyses, providing a foundation for future work on understanding and improving the sociocultural and structural properties of LLM outputs.

Looking ahead, we plan to extend the corpus to additional languages, models, and narrative variables, further broadening the scope for cross-lingual and cross-model analyses. A first line of research we aim to pursue with the corpus is a systematic comparison between LLM-generated and human-written synopses, to better understand potential representation imbalance in machine-generated storytelling.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. Storium: A dataset and evaluation platform for machine-in-the-loop story generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Story generation dataset and platform.

Stefan Baack. 2024. A critical analysis of the largest source for generative ai training data: Common crawl. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2199–2208.

Julia Barroso da Silveira and Ellen Alves Lima. 2024. Racial biases in ais and gemini's inability to write narratives about black people. *Emerging Media*, 2(2):277–287.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Jerome Bruner. 2010. Narrative, culture, and mind. *Telling Stories: Language, Narrative, and Social Life*, 46.

Peter Cardon, Carolin Fleischmann, Jolanta Aritz, Minna Logemann, and Jeanette Heidewald. 2023. The challenges and opportunities of ai-assisted writing: Developing ai literacy for the ai age. *Business and Professional Communication Quarterly*, 86(3):257–295.

Kennedy Casey, Kylee Novick, and Stella F Lourenco. 2021. Sixty years of gender representation in children's books: Conditions associated with overrepresentation of male versus female protagonists. *Plos one*, 16(12):e0260566.

Marc Cavazza. 2025. Large language models preserve semantic isotopies in story continuations. *arXiv preprint arXiv:2510.04400*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*.

Mircea Eliade. 1961. *Mythes, rêves et mystères*. Gallimard.

Jonathan Gottschall. 2012. *The Storytelling Animal: How Stories Make Us Human*. Houghton Mifflin Harcourt.

Geethika Gunti and M Supriya. 2025. Uncovering hidden narratives: Discovering and classifying archetypes in tiny stories. In *2025 International Conference on Advanced Computing Technologies (ICoACT)*, pages 1–6. IEEE.

Sarah Harmon and Sophia Rutman. 2023. Prompt engineering for narrative choice generation. In *International Conference on Interactive Digital Storytelling*, pages 208–225. Springer.

Xi Yu Huang, Krishnapriya Vishnubhotla, and Frank Rudzicz. 2024. The gpt-writingprompts dataset: A comparative analysis of character portrayal in short stories. *arXiv preprint arXiv:2406.16767*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Lucy Li and David Bamman. 2021. Gender and representation bias in gpt-3 generated stories. In *Proceedings of the 3rd Workshop on Narrative Understanding*, pages 48–55.

Helen Lovatt. 2013. *The Epic Gaze: Vision, Gender and Narrative in Ancient Epic*. Cambridge University Press.

Nick Montfort and Rafael Pérez y Pérez. 2023. Computational models for understanding narrative. *Revista de Comunicação e Linguagens*, 58:97–117.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, CA. Association for Computational Linguistics.

Mihai Nadas, Laura Diosan, Andrei Piscoran, and Andreea Tomescu. 2025. Tf1-en-3m: Three million synthetic moral fables for training small, open language models. *arXiv preprint arXiv:2504.20605*.

Partha Pratim Ray. 2023. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3:121–154.

Jill Walker Rettberg. 2024. How generative ai endangers cultural narratives. *Issues in Science and Technology*, 40(2):77–79.

Alexey Tikhonov, Igor Samenko, and Ivan P. Yamshchikov. 2021. Storydb: Broad multi-language narrative dataset. *Computing Research Repository*. ArXiv:2109.14396.