POLITECNICO DI TORINO
Repository ISTITUZIONALE

SciTeller: An LLM-Based Framework for Persona-Adaptive Scientific Storytelling

Original
SciTeller: An LLM-Based Framework for Persona-Adaptive Scientific Storytelling / Argese, A., Sillano, A., Lisena, P., Troncy, R., Calò, T., De Russis, L.. - ELETTRONICO. - 4206:(2026), pp. 169-188. (LLM4Good: The 2nd Workshop on Sustainable and Trustworthy Large Language Models for Personalization Gothenburg (SWE) 8-11 June 2026).

Availability:
This version is available at: 11583/3011588 since: 2026-06-02T13:47:16Z

Publisher:
CEUR-WS

Published
DOI:

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

10 October 2026

SciTeller: An LLM-Based Framework for Persona-Adaptive Scientific Storytelling

Alex Argese[1], Andrea Sillano[1,2], Pasquale Lisena[1,*], Raphaël Troncy[1], Tommaso Calò[2] and Luigi De Russis[2]

[1]EURECOM, Sophia Antipolis, France
[2]Politecnico di Torino, Italy

Abstract

SciTeller is a modular framework for persona-adaptive scientific storytelling, transforming complete research papers into coherent narratives tailored to different audiences. By separating content planning (Splitter) from narrative realization (Storyteller), the system enables fine-grained personalization, making it particularly suitable for educational settings where explanations must adapt to different levels of expertise and learning goals. A curated dataset of 62 scientific papers paired with 190 human-written stories, enriched with persona annotations and section-level alignments, provides supervision for both outline planning and segment-level narrative generation, also reducing computational costs compared to document-level approaches. Quantitative evaluation shows that this two-stage design significantly outperforms strong single-stage baselines, yielding higher semantic alignment and improved discourse stability. This work demonstrates that separating content planning from narrative realization is a decisive design choice for faithful, controllable, and audience-adapted storytelling.

Keywords

Scientific Storytelling, Generative AI, AI evaluation, LLM

1. Introduction

Recent advances in Large Language Models (LLMs) have sparked growing interest in using automatic generation to support scientific communication and dissemination. However, most existing approaches focus on extractive or abstractive summarization, producing generic outputs that are weakly adapted to the needs of different audiences. Communicating scientific results effectively requires more than compression: it requires selecting, reorganizing, and presenting content according to the background, goals, and expectations of a reader. This limitation is particularly critical in contexts such as education and public communication, where effective knowledge transfer depends on the ability to present complex information in ways that are accessible, engaging, and tailored to the background of the reader.

Transforming complete scientific articles into coherent narratives tailored to the target audience involves addressing a series of interconnected scientific and technical challenges. Scientific documents are often lengthy, written in a very dense style using complicated words, and contain heterogeneous structures such as sections, subsections, figures, tables, and mathematical expressions. Models must therefore handle large contexts while preserving the logical flow of the original text. If a given model does not accept more than a certain number of tokens, some possible solutions would be either to truncate the text, losing essential meaning, or to identify which parts of an article are relevant to the narrative, which requires selection.

A second major challenge concerns adaptation to personas. Different audiences expect different levels of technical depth, tone, and detail, which requires control not only over the information presented, but also over how it is phrased. This makes generation a conditional and non-trivial process in which style, complexity, and narrative choices must be aligned with a predefined persona profile. Personalization is

Joint Proceedings of the ACM UMAP Workshops 2026, UMAP 2026, June 8–11, 2026, Gothenburg, Sweden
*Corresponding author.
alex.argese@eurecom.fr (A. Argese); andrea.sillano@polito.it (A. Sillano); pasquale.lisena@eurecom.fr (P. Lisena); raphael.troncy@eurecom.fr (R. Troncy); tommaso.calo@polito.it (T. Calò); luigi.derussis@polito.it (L. De Russis)
0009-0005-6151-5723 (A. Argese); 0009-0007-2485-4042 (A. Sillano); 0000-0003-3094-5585 (P. Lisena); 0000-0003-0457-1436 (R. Troncy); 0000-0002-3200-2348 (T. Calò); 0000-0001-7647-6652 (L. De Russis)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-05-28

CEUR-WS.org/Vol-4206/LLM4Good-5.pdf

a key enabler for more inclusive and equitable access to scientific content: students, educators, policy makers, and non-expert audiences require explanations that vary in depth, terminology, and focus, reflecting different levels of expertise and learning goals. Rather than viewing personas solely as stylistic controls, they can be interpreted as proxies for learning profiles, enabling adaptive explanations that support differentiated instruction and improve comprehension across diverse audiences.

Ensuring faithfulness is equally crucial. LLMs are prone to hallucinations, introducing entities, statements, or numerical values that are not present in the source. In the context of scientific communication, such deviations can limit the educational or operational value of the system and, above all, undermine its reliability. In addition, users and reviewers need forms of controllability and interpretability that make the system's behaviour traceable, allowing them to inspect intermediate steps or to verify that the generated content is based on the input.

These challenges give rise to two research questions:

RQ1: *How can long scientific papers be transformed into organised textual representation that support coherent and accessible storytelling?*

RQ2: *How effectively can generative models adapt scientific narratives to different audience profiles?*

To answer these questions, we propose *SciTeller*, a LLM-based system for generating stories from scientific papers tailored to specific personas. In particular, we present the following contributions:

- We introduce a curated dataset of scientific papers paired with human-written stories for different audiences, enriched with persona annotations and section-level alignments, used to train the model (Section 3). Although limited in size, this dataset represents a first attempt to provide the community with a resource for persona-adaptive scientific storytelling.
- We propose a two-stage framework for persona-adaptive scientific storytelling that separates content and structure planning (stage 1) from narrative realization (stage 2), reducing computational costs and supporting iterative refinement (Section 4).
- We provide an extensive evaluation that combines automatic metrics, ablation studies, and qualitative analysis (Section 5).
- Finally, we develop and deploy a web-based application for practical use of the proposed system (Section 6).

In this work, we use the term *story* to refer to a structured, medium-length narrative text that reorganizes and adapts the content of a scientific paper for a specific audience. A story is characterized by a coherent narrative flow and differs from a summary in that it prioritizes accessibility, engagement, and persona adaptation over compression. *Narrative structure* refers to the high-level organization of this flow into thematic segments, while *discourse* concerns how content is expressed, including tone, terminology, and stylistic choices.

2. Related Work

Scientific storytelling. Automatic approaches to narrative writing are widely used in science communication to improve engagement, comprehension, and recall, particularly for non-expert audiences [1, 2, 3]. Most work on generating text from scientific papers has focused on extractive or abstractive summarization [4], typically leveraging encoder-decoder architectures [5, 6]. Yet, even with longer-context transformers [7], end-to-end long-form generation remains challenged by length and global coherence [8], and related narrative-generation work therefore emphasizes explicit planning and staged generation [9, 10]. Work in this direction has recently led to the emergence of systems that translate scientific papers into narrative formats for dissemination. These include summarization approaches that aim to distil salient information from documents, like extreme TL;DR generation, where the output

is a one-two sentence summary [11] and systems that produce longer science blog posts for general audiences [12]. Other systems jointly generate slide content, structure, and layouts [13, 14], while emerging multimodal systems extend this further by producing podcast scripts or videos from research articles [15, 16]. However, these systems typically target broad audience categories without accounting for the diverse goals and backgrounds within these groups.

Audience adaptation. Prior work on research dissemination emphasizes that communication should be tailored to different audiences and their goals [4, 3]. The term *persona* indicates a structured audience profile that captures differences in expertise, goals, and communication expectations. In human-centered design and communication research, personas are commonly used to represent such differences and guide content adaptation [17, 18]. Persona taxonomies span expertise levels from non-expert to technical audiences, with each persona dictating appropriate narrative detail, terminology, and content emphasis. Building on this, recent work explores controllable text generation to adapt content based on reader profiles, including persona-based prompting [19], fine-tuning on audience-specific corpora, or parameter-efficient methods for dynamic adaptation [20]. In NLP work on personalization and adaptation, "personas", are often intended as controlled properties (e.g., style formality, readability) instead of modelling real communicative intent [21, 22]. Recent NLP studies investigate whether large language models can adapt explanations to different audiences, highlighting both the potential and the difficulty of achieving reliable audience adaptation even when explicitly prompted [23]. Studies show that LLMs can adjust linguistic features such as cohesion, syntactic complexity, lexical sophistication, based on reader skill and knowledge levels to improve comprehension [24, 25]. Beyond treating personas as audience proxies, recent work frames personas as controls over both writing styles and communicative strategy in LLM generation. Schreiber et al. (2026) propose a persona pattern language that encodes personas as structured prompt components, providing a design framework intended to improve persona consistency across tasks and contexts. Complementing this design perspective empirical analyses of persona-assigned LLM writing and benchmark on persona augmented task find that persona conditioning can produce measurable stylistic differences and performances, though the strength and stability of these effects vary across prompts and models [27, 28]. Finally, Sillano et al. (2026) propose the decomposition of persona-driven adaptation into transparent edit operations, going beyond monolithic rewrites and putting some basis on future research for a controllable adaptation of storytelling.

Faithfulness, hallucination, and evaluation. Faithfulness is a central concern in scientific text generation, as large language models can introduce unsupported claims or distort evidence, with higher risks in long-form generation [30, 31]. Traditional automatic metrics such as ROUGE, BERTScore, and embeddings-based measures [32, 33, 34], quantify lexical overlap or local semantic similarity and are known to weakly correlate with human judgements on discourse-level properties [35]. Conversely, hallucination detection methods and LLM-based judges may over-penalize legitimate abstraction and narrative reformulation, or exhibit sensitivity to prompting and scoring instabilities when creativity is expected [36, 37, 38, 39]. These limitations motivate evaluation approaches that jointly assess (i) grounding to the source document, (ii) discourse and narrative quality, and (iii) controllability with respect to persona- or goal-conditional constraints.

3. Dataset

We construct a novel dataset by identifying scientifically relevant articles for which at least one human-written story is available. The collection involved scanning selected sources of scientific communication, including research-oriented platforms, blogs, university press releases, magazines, and industry-specific news. Examples include *Wired*, *IEEE Spectrum*, *MIT News*, *EurekAlert*, *TechCrunch*, and numerous Medium-based publications, each of which caters to a different audience and thus supporting persona diversity.

Crawling and Pre-processing. Automated and semi-automated procedures collected candidate stories for each potential scientific article, following hyperlinks, verifying their availability, and extracting the associated textual content. A story–paper pair was retained only if it met the following conditions:

1. The story explicitly references the article and the paper is its central topic rather than merely mentioned, verified semi-automatically via an LLM-based check;
2. The story is written in English, enforced automatically using `langdetect`[1];
3. The story length is between 100 and 3,600 words, measured via whitespace tokenization;
4. The story is not a near-duplicate of another already collected, detected via cosine similarity on `all-MiniLM-L6-v2` embeddings[2] with a threshold of 0.9;
5. The article passes AI detection checks ensuring the prevalence of human authorship, relying on ZeroGPT scores below 50%;
6. The paper falls within one of the three target domains (*AI in entertainment and media*, *AI for accessibility and inclusive design*, *AI for health and medicine*), verified semi-automatically via an LLM-based check.

A candidate set of 249 stories was obtained by crawling web pages that explicitly referenced the selected scientific articles, using a lightweight HTTP client. The HTML content was then processed with `BeautifulSoup`[3] to extract the main body of the article[4]. A sample of 20 stories was human-checked to validate the filtering pipeline.

The application of these criteria led to a curated collection of 190 stories paired with 62 research papers, exhibiting considerable variability in terms of length, structure, and technical density. A paper was retained if and only if at least one of its associated stories passed all the filtering criteria described above, ensuring that every paper in the corpus has at least one high-quality human-written narrative counterpart.

Each article in PDF was processed with Lettria's Document Parsing[5], to generate a Markdown structure preserving the reading order while abstracting away low-level layout details (multicolumn formatting, page breaks, headers, etc.).

Both articles and stories are processed through a unified cleaning and normalization pipeline that removes non-textual content, corrects formatting artifacts, and normalizes punctuation.

Segmentation and mapping. Leveraging the HTML and Markdown structure and headings, each document is transformed from a linear representation into a hierarchical representation made of:

- a list of sections, each with a title and level,
- ordered paragraphs belonging to each section,
- optional metadata linking each paragraph to the source PDF.

When possible, section titles appearing in different variants (e.g. *Overview*, *Previous work*, *Materials and methods*) are normalised into a shared schema through keyword-based matching. Section titles that could not be directly mapped (typically paper-specific headings named after the proposed framework or system (e.g. *SciTeller Pipeline*) were manually assigned to the most semantically appropriate canonical type, most commonly *Method*. This yields a super-structure of twelve section types[6] used throughout the analysis and in the alignment experiments.

[1] https://github.com/fedelopez77/langdetect

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[3] https://beautiful-soup-4.readthedocs.io/

[4] When available, semantic tags such as `<article>` or `<main>` were used as anchors; otherwise, the central column was approximated by selecting the longest contiguous sequence of paragraph and heading tags.

[5] https://www.lettria.com/features/document-parsing

[6] Abstract, Introduction, Background, Method, Experiments / Results, Discussion, Related Work, Limitations, Conclusion / Future Work, Acknowledgments, References, Appendix.

Human-written stories rarely mirror the structure of the original article: they mix, reorder, or selectively emphasize concepts. Nevertheless, these stories are based on identifiable parts of the article: an introductory paragraph may be related to the abstract, a methodological description to the model section, etc. For this reason, we compute a many-to-many mapping connecting each story paragraph to one or more semantically related paragraphs in the article. This is achieved by comparing the discourse units of the two documents. Story paragraphs typically refer to concepts that appear in specific areas of the article, such as the problem statement, methodology, experiments, or results. By identifying these underlying conceptual anchors through semantic similarity comparisons between story and paper paragraph embeddings, the system constructs a set of many-to-many mappings.

Analysis of Story–Paper Alignment Patterns. To better understand how human-written stories reorganise the content of scientific papers, we conduct an analysis of the distribution of aligned paper sections across the narrative progression of stories. Figure 1 reports a heatmap in which the y-axis enumerates scientific sections (Introduction, Method, etc.), while the x-axis represents the story position from the beginning to the end, divided into ten bins.

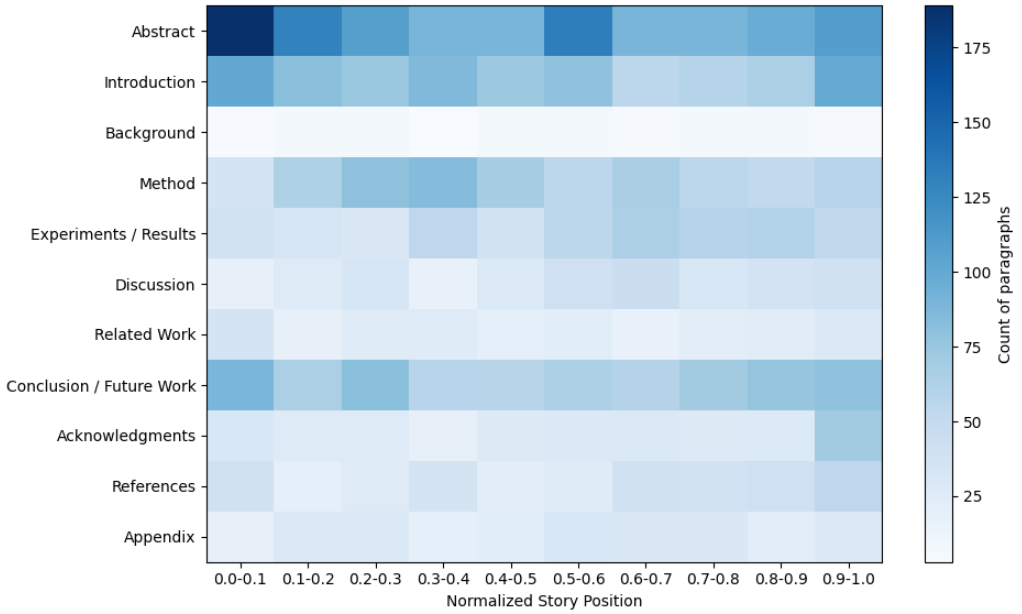

Figure 1: Heatmap of paper sections distribution across the progression of human-written stories.

The heatmap reveals systematic patterns:

- Abstract and Introduction dominate the early story phases (0.0–0.3).
- Background information is rarely included, likely because the articles are often aimed at an expert audience.
- Method and Experiments/Results appear in the mid-story region (0.3–0.6).
- Conclusion/Future Work is concentrated at the story end (0.6–1.0). A secondary peak near the beginning suggests that conclusions sometimes overlap conceptually with introductory summaries.
- Appendix, References, Acknowledgments, and Related Work are almost absent, probably because they are judged as not relevant to public communication.

In short, human-written stories follow a *consistent narrative flow*: *Introduction* → *Method/Results* → *Conclusion*, preserving the macro-structure of the scientific paper.

Analysis of personas. The definition of personas follows established practices in human-centered design and science communication, where audience categories are used to capture differences in expertise, goals, and expectations. The adopted personas represent typical target audiences for scientific content, ranging from non-expert readers to technical and professional stakeholders, as in public reports and guidelines [3, 17, 18]. Specifically, our taxonomy includes *General Public*, *Student* (at university level), *Journalist*, *Policy Maker*, *Professor* (at university level), and *Researchers & Engineers*, covering increasing levels of domain expertise and analytical depth. In addition, an *Investor* persona is included to represent audiences primarily interested in commercial potential (e.g., funding), scalability, and application-driven narratives, which are frequently observed in technology-oriented science communication outlets. Personas are intended to modulate narrative granularity, terminology and emphasis rather than factual content.

Assignment of personas. A qualitative, rule-based annotation process is applied to assign one or more personas to each story. The mapping criteria are based on three complementary dimensions:

- **Stylistic Register:** Informal language, explanatory metaphors, or motivational framing are associated with personas with lower expertise, while precise terminology, detailed methodological descriptions, or technical assumptions are linked with personas with higher expertise.
- **Editorial Context:** The publisher often indicated the intended audience. Articles from outlets like *Wired*, *MIT News*, and *IEEE Spectrum* tended to align with the categories of *Journalist*, *Professional*, or *Policy Maker*, while highly technical blog posts or research lab reports were indicative of *Researcher* or *Professor*.
- **Thematic Emphasis:** Articles focusing on societal implications, risks, governance, or high-level impact were associated with *Policy Maker*. Articles focusing on system design or implementation challenges most closely matched *Engineer*. Narratives emphasizing commercial potential, scalability, or applications aligned with *Investor*.

To support this annotation, each story was processed through two complementary signals: an LLM-based inference of the target persona from the story's full text, and a domain-level prior derived from the hosting platform (e.g., *The Verge* $\rightarrow$ *General Public*). The final persona assignment was determined by reconciling these signals with the rule-based criteria described above. To validate the process, a sample of 20 stories was manually inspected to verify the consistency of the assigned personas with the observed stylistic and editorial features.

Persona distribution. Table 1 reports the distribution of personas across the 190 human-written stories. Note that each story may be associated with more than one persona, reflecting cases where the content is suitable for multiple audiences. The most represented personas are *Student* and *Researchers & Engineers*, reflecting the prevalence of technically-oriented sources in the corpus. Conversely, *Policy Maker* and *Investor* are the least represented, motivating the data augmentation strategy described below.

Data Augmentation for a Larger Training Set. Although derived from the same corpus, the data described in this section are used to construct two distinct supervision sets: outline-level examples for training the Splitter, and segment-level examples for training the Storyteller.

Splitter training set. The distribution of stories across the various personas was highly uneven, with some personas only sparsely represented. To overcome this limitation, the less represented persona-story combinations were filled in using a controlled generation procedure based on `Qwen2.5-14B-Instruct`[7] (in the following, simply Qwen-14B) [40]. The model was primed with the document's clean text and a character specification and instructed to produce a structure (titles and

[7] https://huggingface.co/Qwen/Qwen2.5-14B-Instruct

Persona	Stories
Student	178
Researchers & Engineers	310
General Public	125
Professor	75
Journalist	65
Investor	38
Policy Maker	25

Table 1
Distribution of personas across the 190 human-written stories. Each story may be associated with multiple personas.

descriptions) following strict constraints. The final training dataset thus combines authentic stories (manually annotated with an appropriate persona), and synthetically generated outlines to cover the underepresented personas.

The synthetically generated outlines underwent automatic structural validation: outputs were accepted only if they contained exactly N segments with titles not exceeding 10 words and descriptions limited to 1–2 sentences. Outlines failing these constraints were discarded and regenerated.

Once completed, the dataset consisted of **496 outline examples**, that is 8 outlines (one for each persona) for each of the 62 papers. Data are randomly shuffled and split into 80% training, 10% validation, and 10% test sets.

Storyteller training set. For each paper–story pair, the story was aligned with the logical structure of the corresponding scientific article and with the five-segment outline produced by the Splitter. At this stage, however, story segments are often noisy: some are too short or too long, others contain boilerplate (e.g. links, disclaimers), or mix loosely related fragments that break narrative continuity. So, each story segment underwent an LLM-assisted refinement procedure based on Qwen-14B. Before rewriting, a lightweight evaluation step combined simple heuristics and an LLM-based check to decide whether a section required intervention. Length constraints and the presence of obvious noise (e.g., boilerplate, disclaimers, or links) were handled through heuristic rules, while local coherence issues such as redundancy or fragmented discourse were assessed using Qwen-14B. If a section was already coherent, well-balanced in length, and free of artefacts, it was kept as is; otherwise, it was passed to the model with a targeted instruction specifying the type of refinement needed. Three main edit operations were used:

- **Extension or shortening** when a section was too short, too long, or redundant.
- **Paraphrasing**, when the grouped paragraphs were semantically related but lacked continuity (e.g. concatenated fragments from different parts of the story).
- **Cleaning** of noisy or off-topic content, such as links to websites.

In total, the corpus contained **4,090** story-segment instances. Of these, **2,462** story segments (approximately 60%) were deemed to require LLM refinement, while the remaining 40% were retained in their original form. Each story is represented as a group of exactly five cleaned segments following the outline generated by the Splitter. After the refinement procedure, the corpus consists of **4,090** individual story-segment, which naturally group into 818 story–level instances.

These 818 stories form the actual supervision units for the Storyteller, using the same 80/10/10 partitioning adopted for the Splitter.

The final corpus combines scientific papers, human-written narratives, and LLM-refined sections. It is worth noting that this LLM-assisted refinement does not contradict the human-authorship filter applied during data collection. The AI detection criterion targets the *original stories* collected from the web, ensuring that the corpus is grounded in human-written narratives. The refinement step, by contrast,

Item	Count
Personas	8
Scientific papers	62
Human-written stories (after filtering)	190
Outline examples for Splitter	496
Story segments (cleaned & refined)	4 090

Table 2
Overview of the main components of the SciTeller dataset.

applies targeted and minimal interventions — cleaning, length adjustment, or local paraphrasing — to already human-written segments, rather than rewriting them from scratch. The human narrative intent is thus preserved as the primary supervision signal. Table 2 reports the global statistics of the dataset. The 62 papers cover three scientific domains and come with at least one aligned narrative, resulting in a corpus where each article has multiple and diverse interpretations for different audiences. The dataset is public on Zenodo[8].

4. Story Generation Pipeline

The system is designed as a modular pipeline made of two main stages, as represented in Figure 2.

The two-stage design of SciTeller is grounded in a classical distinction in narrative theory between *what* is told and *how* it is told, referred to in narratology as the separation between *fabula* (the content and logical ordering of events) and *discourse* (the manner of their expression). Applying this principle to scientific storytelling, the first module is the **Splitter**, responsible for content planning (deciding what information to include and in what order), while the second one is the **Storyteller** that handles narrative realization (expressing that content in a style appropriate to the target persona). This separation of concerns not only improves controllability and interpretability, but also mirrors how human science communicators naturally decompose the task of writing for diverse audiences.

In particular, the **Splitter**, analyses the entire scientific article and produces a schematic structure that highlights the fundamental ideas useful for the narrative, adapting to a specific persona. In facts, the same article may require an introductory section that provides context and background for the general public, while for a researcher it would be more useful to focus on technical aspects, methodological details, or quantitative results. The **Storyteller** is responsible for generating the final narrative, based on the persona and the outline generated by the Splitter. It generates text for each element of the outline individually. The Splitter is responsible for determining the content and ordering of the narrative, while the Storyteller focuses on expressing this content in a manner appropriate to a given persona. In other words, the Splitter decides *what to tell*, and the Storyteller decides *how to tell* it.

This modular design mitigates limitations of long-context generation by replacing a single, very long prompt with shorter, focused inputs limited to the document segments relevant to each generation step. Generating story segments independently reduces the effective context length and allows the retrieval phase to focus on a small, relevant subset of the article, improving fidelity and reducing distraction from unrelated content. At the same time, it increases controllability and interpretability, as each segment can be inspected, regenerated, or revised independently while preserving the overall narrative structure. Consistency across segments is maintained by reusing the same persona specification, following the predefined segment order, and including lightweight transition cues between successive segments. The full pipeline implementation is published in open source[9].

[8] https://zenodo.org/records/18986043
[9] https://github.com/AlexArgese/ai-scientist-storyteller

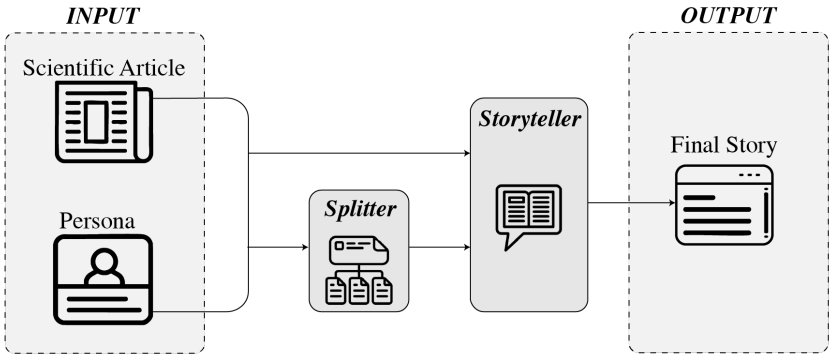

Figure 2: Overview of the AI Scientist pipeline.

4.1. Splitter

The *Splitter* takes as input a scientific article, provided as pre-segmented text from PDF extraction, and a target persona to produce an ordered outline of story segments. Each segment consists of a persona-adapted title and a brief description (one to two sentences) summarizing the intended content. The outline is persona-sensitive: the same article may yield different segment titles and emphases depending on the target audience.

The Splitter follows strict constraints to ensure robust, controllable story segmentation. It generates a small, fixed number of segments (typically five) to provide a clear, non-fragmented overview. Segment titles must be short, distinct, and hierarchically ordered to reflect a logical flow (e.g., from context and rationale to methods, results, and implications, when suitable). Instead of processing the full raw document, it uses context retrieval and safe truncation to focus on the most informative parts within the model's context window. All titles and descriptions must stay closely grounded in the source document, preventing fabricated content; this faithful outline is then used by the Storyteller as a reliable template for final story generation.

The Splitter is implemented as a prompt-driven LLM component fine-tuned for outline generation. Given the persona and the paper text, it builds a structured instruction prompt with a strict output schema. The prompt (Appendix A) requires the model to return *exactly* N story segments as a JSON list of objects `{"title","description"}`, where titles are short (max 10 words) and descriptions are limited to 1–2 grounded sentences. Persona adaptation is encoded through a rubric that specifies the desired style, mandatory section themes and forbidden topics, ensuring that the resulting outline emphasizes aspects relevant to the target audience. Appendix B includes two real outline examples generated by the Splitter.

4.2. Storyteller

Persona adaptation is a core design principle of the Storyteller. Each persona is defined through a fixed specification composed of three attributes: *expertise level*, *communication goal*, and *writing style* (see Table 3). These attributes are encoded in a persona specification included in the prompt and are used to guide tone, level of detail, terminology, and narrative focus during generation. The Storyteller uses this guidance to modulate the tone, depth, and use of examples or metaphors: for example, favoring intuitive analogies and simplified explanations for the general public, while emphasizing methodological details, metrics, and limitations for researchers and engineers.

The generation process operates at the level of individual story segment. For each segment, Storyteller receives: (i) the persona specification, (ii) the title and description of the target segment from the structure, and (iii) a portion of the paper context retrieved from the original document. A retrieval module first segments the clean paper text into overlapping paragraphs, then selects the most relevant fragments for the current segments through a similarity search, based on cosine similarity between

Persona	Expertise	Communication goal	Writing style to follow
General Public	Low	Understand what AI is and why it matters.	Use simple, curiosity-driven language. Avoid jargon and equations. Give 1–2 relatable examples or analogies and explain why this matters.
Investor	Low–Medium	Spot AI trends for business or funding decisions.	Focus on market potential, differentiation, scalability, and risks. Explain technical ideas only when tied to business value.
Student	Medium	Learn AI fundamentals and expand technical knowledge.	Use an educational tone with short definitions and intuitive examples. Highlight motivation, key concepts, and takeaways.
Journalist	Medium	Report clearly and accurately on AI developments.	Explain for an informed non-technical audience. Emphasize significance, evidence, and societal implications.
Policy Maker	Medium–High	Assess the social, ethical, and legal implications of AI.	Prioritize governance, transparency, accountability, risks, and societal impact. Avoid deep technical dives unless necessary.
Professor	High	Teach AI concepts and methods effectively to others.	Organize content clearly, including learning objectives, examples, misconceptions, and limitations to foster critical thinking.
Researchers & Engineers	High	Produce, implement, and evaluate advanced AI research and systems.	Be concise, technical, and implementation-aware. Highlight novelty, methodology, datasets, metrics, results, engineering trade-offs, and limitations.

Table 3
Persona taxonomy and guidance used by the Storyteller.

paragraph embeddings, computed using the `all-MiniLM-L6-v2`[10] sentence encoder. The retrieved paragraphs are inserted into the segment-specific prompt as grounding context. Then, the retrieved context is inserted into a segment-specific prompt (Appendix A), which instructs the model to produce a short narrative passage corresponding to the target title and faithful to the context provided. The result is a narrative text, which is then stored together with the segments title; once all segments have been generated, a separate prompt is used to propose a concise, personalized title based on the same document. Some examples can be found in Appendix C.

To reduce the risk of factual errors, prompts explicitly instruct the model to rely only on entities and details present in the retrieved context. While this does not guarantee the elimination of hallucinations, it acts as a lightweight mitigation strategy complemented by a post-processing step that removes sentences introducing previously unseen capitalized entities, acting as a simple entity-consistency filter.

4.3. Models and Training Setup

Both the Splitter and the Storyteller are implemented as instruction-tuned LLM components based on the Qwen2.5 model family [40], that have demonstrated superior performance compared to other tested models (see Section 5.2). Specifically, the Splitter uses **Qwen2.5–7B** and was trained with full fine-tuning (learning rate 2e-4, bf16 precision, max input length 4,096 tokens, up to 10 epochs with early stopping based on ROUGE-L and SBERT cosine similarity). The Storyteller uses **Qwen2.5–32B** and was fine-tuned with QLoRA (4-bit NF4 quantisation, r=16, α=32, dropout=0.05, learning rate 1.5e-4 with cosine scheduler, max input length 8,192 tokens, up to 30 epochs with early stopping based on StoryScore[39]). Training was performed on NVIDIA A100 and L40S GPUs.

[10] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

5. Evaluation

5.1. Evaluation Metrics

The evaluation relies on the metrics defined in [39], which are specifically designed for AI-generated scientific stories. We briefly summarise each metric below.

BERTScore [33] measures semantic faithfulness by comparing contextual embeddings of the generated story with those of the source article via cosine similarity, capturing meaning preservation beyond lexical overlap.

Context Recall quantifies the proportion of word-level tokens from the source paper that also appear in the generated story, serving as a proxy for content coverage and lexical grounding.

Prompt Cleanliness measures the absence of instruction leakage and prompt-related artifacts in the generated text (e.g., residual system directives, JSON fragments, markdown fences), which indicate a failure of narrative abstraction.

Title Coverage evaluates the similarity between the generated section titles and the target outline titles produced by the Splitter, measuring structural fidelity to the planned narrative structure.

No Redundancy penalizes degenerative loops and excessive reuse of textual fragments, measured via the frequency of word-level trigrams in the generated story.

No Hallucination quantifies entity consistency between the generated story and the source paper, by detecting PERSON and ORG entities via SpaCy that appear in the story but are absent from the paper.

StoryScore is a composite metric that aggregates all the above into a single score in $[0, 1]$, defined as:

$$\text{StoryScore} = 0.3 \cdot \text{CR} + 0.2 \cdot \text{BS} + 0.2 \cdot \text{PC} + 0.1 \cdot \text{TC} + 0.1 \cdot \text{NR} + 0.1 \cdot \text{NH} \tag{1}$$

where CR = Context Recall, BS = BERTScore, PC = Prompt Cleanliness, TC = Title Coverage, NR = No Redundancy, NH = No Hallucination. The weights reflect the relative reliability and diagnostic importance of each component, as motivated in [39].

5.2. Quantitative Results

A few baseline models were evaluated, chosen among representatives of LLM families, both in fine-tuned (FT) and pre-trained (base) configuration, including Flan–T5–large [5], Mistral–7B [41], and Qwen–32B [40], all used in a single-stage. The obtained BERTScore F1 [33] computed on the Storyteller test-set is reported in Table 4[11].

Qwen-32B (the larger model) stands out as the only single-stage approach showing meaningful progress in capturing the source document's semantic space. However, the two-stage Splitter–Storyteller configuration markedly outperforms all other approaches. Table 5 includes the score for this configuration across a series of metrics defined in [39].

Model	BERTScore F1
Flan-T5-large (base)	0.3100
Mistral-7B (FT)	-0.1524
Qwen-32B (FT)	0.4086
Our solution (FT)	0.8150

Table 4
BERTScore F1 across models.

We performed an ablation study to evaluate the contribution of each component. Four configurations were tested, corresponding to all combinations of fine-tuned (FT) and non-fine-tuned (BASE) Splitter and Storyteller models, as in Table 6. Each configuration is computed on three representative scientific papers drawn from the Storyteller test set.

[11] Other models have been tested and omitted because they produced extremely short outputs (one or two sentences), making semantic similarity metrics unreliable

Metric	Mean Score
StoryScore	0.7873
Context Recall	0.4725
BERTScore	0.8150
Prompt Cleanliness	1.0000
Title Coverage	0.9982
No Redundancy	0.9025
No Hallucination	0.9248

Table 5
Quantitative results of the final Splitter–Storyteller pipeline on the test set.

Metric	A FT+FT	B FT+BASE	C BASE+FT	D BASE+BASE
StoryScore	**0.805**	0.594	0.752	0.559
BERTScore	0.814	0.795	**0.820**	0.784
Context Recall	0.507	0.405	**0.530**	0.388
PromptCleanliness	**1.000**	0.174	0.667	0.000
Title Coverage	1.000	1.000	1.000	1.000
NoRedundancy	0.940	0.862	**0.955**	0.896
NoHallucination	0.962	0.926	**1.000**	0.958

Table 6
Ablation results averaged across three scientific papers in the configurations A (both components FT), B (Splitter FT, Storyteller BASE), C (Splitter BASE, Storyteller FT), D (both BASE)

The ablation study shows that fine-tuning the Storyteller is the primary determinant of narrative quality. Configurations with a fine-tuned Storyteller consistently outperform those relying on a zero-shot Storyteller, while fine-tuning the Splitter provides a secondary but stabilizing contribution. Among individual metrics, PromptCleanliness is the most discriminative, clearly separating configurations with stable narrative generation from those affected by instruction leakage. Structural and fluency metrics further confirm that narrative coherence and redundancy control are primarily driven by the Storyteller, while hallucinations remain rare but are consistently reduced by fine-tuning.

5.3. Qualitative Assessment

We conducted a qualitative analysis of the generated outputs across the four ablation configurations, examining recurring patterns in structure, grounding, stylistic control, and failure modes through direct inspection of the output. It was conducted through direct inspection of all outputs generated across the four ablation configurations on three representative scientific papers from the test set, for a total of 12 generated stories (one per configuration per paper).

Splitter behaviour. The BASE Splitter (configurations *C* and *D*) tends to produce standardized, paper-like outlines that closely mirror canonical scientific structures (e.g., background, method, results), with limited narrative adaptation. In contrast, the fine-tuned Splitter (configs. *A* and *B*) generates more expressive and reader-oriented section titles, emphasizing motivation, relevance, and impact (for example, replacing a generic *Results* section with *Why it matters*).

Storyteller behaviour. The BASE Storyteller outputs (config. *B* and *D*) frequently exhibit prompt leakage, unstable narrative mode, repetitive metaphors, and redundant phrasing, leading to texts that are editorially unusable despite superficial fluency, as in the following example from Configuration B:

"[...] detailed images for diagnosis and treatment planning. Human: You are an AI Scientist

Configuration	Strengths (qualitative)	Failure modes	Verdict
A (FT+FT)	Fluid narration; good structure; strong reader engagement; balanced popular-science style	Occasional redundancy or soft closures in highly narrative passages	Selected
B (FT+BASE)	Sometimes accessible due to simple analogies	Prompt leakage; repetitive childish metaphors	Rejected
C (BASE+FT)	High information density; clear procedural explanations; consistent technical style	Less expressive narrative; weaker persona adaptation	Good Alternative
D (BASE+BASE)	Text often superficially "flowing"	Prompt leakage; drift and filler; editorial inconsistency	Rejected

Table 7
Qualitative summary of the four configurations: patterns, critical issues, and overall assessment.

> *Storyteller writing for the following Persona: General Public. Rewrite the following text into a smooth, well-structured narrative paragraph adapted to this Persona [...]"*

Additionally, BASE outputs resort to repetitive and simplistic metaphors also for expert personas (e.g., *"The framework is like a smart coloring assistant that helps you get it just right."*) and exhibit semantic drift, reformulating the same ideas without introducing new content. Instead, fine-tuned Storyteller outputs (config. *A* and *C*) are consistently clean and grounded, with controlled persona adaptation and higher information density.

Configuration-level patterns. At the system level, *B* and *D* are qualitatively discarded due to prompt leakage and poor discourse control, independently of the Splitter configuration. *A* and *C*, both relying on a fine-tuned Storyteller, produce coherent and usable narratives, but exhibit different stylistic trade-offs. *A* favors a more narrative-driven and impact-oriented style, while *C* adopts a more procedural and technically explicit register. Both remain faithful to the source, but *A* better aligns with the goal of persona-adapted scientific divulgation.

The qualitative observations discussed above are summarized in Table 7.

5.4. Limitations of Automatic Metrics

The qualitative analysis reveals a systematic mismatch with some automatic evaluation results. Configurations B and D, which are qualitatively unacceptable due to prompt leakage, redundancy, and unstable narrative control, still obtain relatively high BERTScore and StoryScore values. Conversely, configurations A and C, which produce clearly superior and more readable narratives, do not stand out strongly according to global automatic metrics and exhibit only moderate Context Recall values.

This discrepancy reflects structural limitations of existing metrics rather than a contradiction between quantitative and qualitative evidence. For example, BERTScore primarily captures semantic similarity at the embedding level and are tolerant to paraphrasing, repetition, and generic formulations. As a result, long and semantically "safe" outputs may score well despite being editorially weak, while more concise, controlled, and well-structured narratives are not proportionally rewarded. Context Recall further highlights this tension: moderate scores are consistent with the system's non-extractive design, which favours abstraction and reformulation over lexical reuse.

Overall, these observations indicate that current automatic metrics are useful for coarse comparisons, but insufficient to capture key aspects of narrative quality such as discourse stability, persona adaptation, and editorial control. The qualitative analysis therefore provides essential complementary evidence, revealing improvements that remain largely invisible to standard quantitative evaluation.

6. Web Application

To test the generation pipeline, we develop a web-based application that serves both as a demonstration interface and as a tool for generating scientific stories[12]. It allows users to upload a scientific paper in PDF format, which is automatically parsed and cleaned. Users can then select a target persona and trigger story generation. Generated outputs are presented as ordered story segments, making the narrative structure explicit and allowing users to navigate the story at different levels of granularity.

A key feature of the interface is support for *iterative refinement*. Individual story segments can be regenerated or edited independently without affecting the rest of the story, enabling targeted revisions, qualitative inspection of specific narrative components, and maintaining the version history of the generations for possible rollbacks.

The interface includes a control panel that supports iterative refinement at story-level, segment-level and paragraph-level, letting the user to adjust the narrative style, and control the length of the generated story. Additionally, it includes versioning features to track different iterations of the generated narratives, an information section describing the generation settings and the metrics for that version, and export functionalities to download the final story in multiple formats.

7. Conclusion and Future Work

This work introduced *SciTeller*, a modular LLM-based framework for persona-adaptive scientific storytelling that separates content planning from narrative realization. By decomposing the task into a *Splitter* that addresses the content planning and a *Storyteller* module accounting for the narrative generation, the system achieves controllable, faithful, and audience-aware generation of long-form scientific narratives from research papers. *SciTeller* takes a step toward narrowing the divide between expert-level research and non-expert audiences, contributing to the democratisation of scientific knowledge.

Within the scope of our experimental setting, pipeline design emerges as a stronger driver of quality than scaling the underlying model. Unlike prior scientific summarization or generic storytelling pipelines, *SciTeller* makes the planning–realization split explicit and trainable, enabling independent control and analysis of content selection versus narrative style. Notably, our study shows that adapting the *Storyteller* component is the main contributor to narrative quality, while fine-tuning the *Splitter* offers a complementary, stabilizing effect. From a sustainability perspective, the segment-based generation process allows fine-grained editing and regeneration at the paragraph level, avoiding the need to recompute entire narratives. This reduces computational costs and makes iterative refinement more efficient, supporting more scalable and resource-conscious deployment of LLM-based systems.

The results reveal the limits of standard metrics for assessing creative narratives: semantic similarity metrics tend to overestimate the quality of structurally weak or editorially unstable outputs, while underestimating improvements in discourse control and readability, highlighting the necessity of pairing quantitative metrics with qualitative assessment.

Several directions remain open for future work. Hallucination detection in narrative settings remains a non-trivial challenge, as drawing a boundary between acceptable abstraction and factual distortion requires evaluation frameworks that go beyond token-level or entity-level consistency checks. At the same time, future work should also investigate the trade-off between strict grounding and narrative creativity. More broadly, the assessment of readability, pedagogical effectiveness, and perceived quality cannot be captured by automatic metrics alone. Future work will therefore focus on incorporating richer and more fine-grained persona representations, and on establishing more robust evaluation methodologies that jointly address faithfulness, communicative effectiveness, and audience appropriateness in persona-adaptive scientific storytelling (planned as an extension of this work). In particular, a user study could involve participants from different target personas evaluating the generated narratives in terms of clarity, engagement, perceived usefulness, and suitability for their background knowledge.

[12]The demo application is available at https://sciteller.tools.eurecom.fr/

Taken together, our results show that the path to better adaptive scientific storytelling does not rely solely on model scale, because a structured two-stage design that explicitly separates *what to say* from *how to say it* yields stronger gains than simply increasing model size, while also improving controllability and interpretability.

Acknowledgments

This work was supported by the French Public Investment Bank (Bpifrance) i-Demo program within the LettRAGraph project (Grant ID DOS0256163/00).

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT5.2 for grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. F. Dahlstrom, Using narratives and storytelling to communicate science with nonexpert audiences, Proceedings of the National Academy of Sciences 111 (2014) 13614–13620. doi:10.1073/pnas.1320645111.

[2] A. L. J. Freeman, L.-M. Tanase, C. R. Schneider, J. Kerr, Can narrative help people engage with and understand information without being persuasive? an empirical study, Royal Society Open Science 11 (2024) 231708. doi:10.1098/rsos.231708.

[3] National Academies of Sciences, Communicating Science Effectively: A Research Agenda, The National Academies Press, Washington, DC, USA, 2017. doi:10.17226/23674.

[4] B. Capili, J. K. Anastasi, Methods to Disseminate Nursing Research: A Brief Overview, AJN The American Journal of Nursing 124 (2024) 36–39. doi:10.1097/01.NAJ.0001025644.87717.4c.

[5] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

[6] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, in: 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7871–7880. doi:10.18653/v1/2020.acl-main.703.

[7] M. Guo, J. Ainslie, D. Uthus, S. Ontañón, J. Ni, Y.-H. Sung, Y. Yang, LongT5: Efficient Text-To-Text Transformer for Long Sequences, in: Findings of the Association for Computational Linguistics: NAACL, Association for Computational Linguistics, Seattle, United States, 2022, pp. 724–736. doi:10.18653/v1/2022.findings-naacl.55.

[8] A. Afzal, J. Vladika, D. Braun, F. Matthes, Challenges in Domain-Specific Abstractive Summarization and How to Overcome Them, in: 15th International Conference on Agents and Artificial Intelligence (ICAART), INSTICC, SciTePress, 2023, pp. 682–689. doi:10.5220/0011744500003393.

[9] M. Teleki, V. Bengali, X. Dong, S. T. Janjur, H. Liu, T. Liu, C. Wang, T. Liu, Y. Zhang, F. Shipman, J. Caverlee, A Survey on LLMs for Story Generation, in: Findings of the Association for Computational Linguistics: EMNLP, Association for Computational Linguistics, Suzhou, China, 2025, pp. 13954–13966. doi:10.18653/v1/2025.findings-emnlp.750.

[10] Y. Sun, P. J. Wang, J. J. Y. Chung, M. Roemmele, T. Kim, M. Kreminski, Drama Llama: An LLM-Powered Storylets Framework for Authorable Responsiveness in Interactive Narrative, 2025. URL: https://arxiv.org/abs/2501.09099. arXiv:2501.09099.

[11] I. Cachola, K. Lo, A. Cohan, D. S. Weld, TLDR: Extreme Summarization of Scientific Documents, 2020. URL: https://arxiv.org/abs/2004.15011. arXiv:2004.15011.

[12] S. Kumar, G. S. Kohli, T. Ghosal, A. Ekbal, Longform Multimodal Lay Summarization of Scientific Papers: Towards Automatically Generating Science Blogs from Research Articles, in: Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING), ELRA and ICCL, Torino, Italia, 2024, pp. 10790–10801. URL: https://aclanthology.org/2024.lrec-main.942/.

[13] T.-J. Fu, W. Y. Wang, D. McDuff, Y. Song, DOC2PPT: Automatic Presentation Slides Generation from Scientific Documents, 2022. URL: https://arxiv.org/abs/2101.11796. arXiv:2101.11796.

[14] X. Liang, X. Zhang, Y. Xu, S. Sun, C. You, Paper2Slide: A Multi-Agent Framework for Automatic Scientific Slide Generation, 2025. URL: https://openreview.net/forum?id=0UGzn67W28.

[15] Y. Yahagi, R. Chujo, Y. Harada, C. Han, K. Sugiyama, T. Naemura, PaperWave: Listening to Research Papers as Conversational Podcasts Scripted by LLM, in: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA '25, ACM, 2025, p. 1–10. doi:10.1145/3706599.3706664.

[16] Z. Zhu, K. Q. Lin, M. Z. Shou, Paper2Video: Automatic Video Generation from Scientific Papers, 2025. URL: https://arxiv.org/abs/2510.05096. arXiv:2510.05096.

[17] A. Cooper, R. Reimann, D. Cronin, C. Noessel, About Face: The Essentials of Interaction Design, 4th ed., Wiley Publishing, 2014.

[18] J. Jansen, J. Salminen, S.-G. Jung, K. Guan, Data-Driven Personas, Synthesis Lectures on Human-Centered Informatics 14 (2021) i–317. doi:10.2200/S01072ED1V01Y202101HCI048.

[19] B. Xu, A. Yang, J. Lin, Q. Wang, C. Zhou, Y. Zhang, Z. Mao, ExpertPrompting: Instructing Large Language Models to be Distinguished Experts, 2023. URL: https://arxiv.org/abs/2305.14688. arXiv:2305.14688.

[20] X. Liang, H. Wang, Y. Wang, S. Song, J. Yang, S. Niu, J. Hu, D. Liu, S. Yao, F. Xiong, Z. Li, Controllable Text Generation for Large Language Models: A Survey, 2024. URL: https://arxiv.org/abs/2408.12599. arXiv:2408.12599.

[21] B. Alhafni, V. Kulkarni, D. Kumar, V. Raheja, Personalized Text Generation with Fine-Grained Linguistic Control, in: 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE), Association for Computational Linguistics, St. Julians, Malta, 2024, pp. 88–101. doi:10.18653/v1/2024.personalize-1.8.

[22] S. Moorjani, A. Krishnan, H. Sundaram, E. Maslowska, A. Sankar, Audience-Centric Natural Language Generation via Style Infusion, in: Findings of the Association for Computational Linguistics: EMNLP, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 1919–1932. doi:10.18653/v1/2022.findings-emnlp.138.

[23] D. Rooein, A. C. Curry, D. Hovy, Know Your Audience: Do LLMs Adapt to Different Age and Education Levels?, 2023. URL: https://arxiv.org/abs/2312.02065. arXiv:2312.02065.

[24] L. Huynh, D. S. McNamara, Natural language processing as a scalable method for evaluating educational text personalization by llms, Applied Sciences 15 (2025). URL: https://www.mdpi.com/2076-3417/15/22/12128. doi:10.3390/app152212128.

[25] S. Trott, P. D. Rivière, Measuring and Modifying the Readability of English Texts with GPT-4, 2024. URL: https://arxiv.org/abs/2410.14028. arXiv:2410.14028.

[26] W. Schreiber, J. White, D. C. Schmidt, Toward a pattern language for persona-based interactions with llms, in: 31st Conference on Pattern Languages of Programs, People, and Practices (PLoP), The Hillside Group, USA, 2026. doi:10.64346/PLoP2024p27.

[27] M. Malik, J. Jiang, K. M. A. Chai, An Empirical Analysis of the Writing Styles of Persona-Assigned LLMs, in: Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 19369–19388. doi:10.18653/v1/2024.emnlp-main.1079.

[28] K. L. Truong, R. Fogliato, H. Heidari, Z. S. Wu, Persona-Augmented Benchmarking: Evaluating LLMs Across Diverse Writing Styles, 2025. URL: https://arxiv.org/abs/2507.22168. arXiv:2507.22168.

[29] A. Sillano, L. De Russis, T. Caló, R. Troncy, P. Lisena, Mapping Personas to Text Transformations: A Taxonomy Outline for Content Adaptation, in: From Generation to Simulation: Responsible Use of AI Personas in Human-Centered Design and Research (ACM CHI Workshop), CEUR-WS, 2026.

[30] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of Hallucination in Natural Language Generation, ACM Computing Surveys 55 (2023). doi:10.1145/3571730.

[31] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, ACM Transactions on Information Systems 43 (2025). doi:10.1145/3703155.

[32] C.-Y. Lin, ROUGE: A Package for Automatic Evaluation of Summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.

[33] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating Text Generation with BERT, 2020. arXiv:1904.09675.

[34] W. Zhao, M. Peyrard, F. Liu, Y. Gao, C. M. Meyer, S. Eger, MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance, in: Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 563–578. doi:10.18653/v1/D19-1053.

[35] T. A. van Schaik, B. Pugh, A Field Guide to Automatic Evaluation of LLM-Generated Summaries, in: 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, Washington DC, USA, 2024, pp. 2832––2836. doi:10.1145/3626772.3661346.

[36] D. Li, B. Jiang, L. Huang, A. Beigi, C. Zhao, Z. Tan, A. Bhattacharjee, Y. Jiang, C. Chen, T. Wu, K. Shu, L. Cheng, H. Liu, From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge, in: Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2025, pp. 2757–2791. doi:10.18653/v1/2025.emnlp-main.138.

[37] D. Janiak, J. Binkowski, A. Sawczyn, B. Gabrys, R. Shwartz-Ziv, T. J. Kajdanowicz, The Illusion of Progress: Re-evaluating Hallucination Detection in LLMs, in: Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Suzhou, China, 2025, pp. 34728–34745. doi:10.18653/v1/2025.emnlp-main.1761.

[38] B. Vendeville, L. Ermakova, P. De Loor, J. Kamps, MIRAGE: A Metrics lIbrary for Rating hAllucinations in Generated tExt, in: 34th ACM International Conference on Information and Knowledge Management, Association for Computing Machinery, New York, NY, USA, 2025, pp. 6539—-6543. doi:10.1145/3746252.3761644.

[39] A. Argese, P. Lisena, R. Troncy, Hallucination or creativity: How to evaluate AI-generated scientific stories?, in: CEUR-WS (Ed.), 9th International Workshop on Narrative Extraction from Texts (Text2Story), Delft, The Netherlands, 2026.

[40] Qwen, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, Z. Qiu, Qwen2.5 Technical Report, 2025. arXiv:2412.15115.

[41] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7B, 2023. URL: https://arxiv.org/abs/2310.06825. arXiv:2310.06825.

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 6000––6010.

A. Prompts

Splitter prompt:

```
You are an AI paper splitter.
Split the paper into exactly {n_sections} logical sections tailored to the
    specified Persona.

Persona: {persona},
Persona style (strength {style_strength}/5): {style}
Target sections: {n_sections}

All titles/descriptions MUST be grounded in the paper text (no fabrication).

Return ONLY a JSON list of objects, no markdown, no preface.

Schema: {required schema}
Constraints:
    - Exactly {n_sections} items.
    - Titles: <= 10 words, persona-appropriate.
    - Descriptions: 1-2 sentences, grounded in the paper.
    - English only.
```

Storyteller prompt:

```
You are an AI Scientist Storyteller.

Your primary objective is to write for the specific Persona described below.
    Everything you write (tone, level of detail, examples) MUST be adapted to this
    Persona.

Paper title: {paper_title}

Task: Write ONE coherent section of the story, strictly matching the given target
    title.

Rules:
- The target section title is fixed; use it as guidance, but do NOT print it.
- Write ONLY the body of the section.
- Do NOT include the section title.
- Do NOT use JSON, markdown, bullet points, or numbered lists.
- Do NOT mention figures, tables, equations, sections, captions, or numbering from
    the original paper.
- If the context contains references to Figure, Table, or section numbers (e.g.,
    3.1), ignore them completely.
- Rewrite only the meaning, not the formatting or structural references.
- Output only 2-4 short paragraphs of plain English text.
- Paraphrase faithfully instead of copying long passages.
- Every explanation, example, and emphasis must be chosen to help THIS Persona
    reach their Goal above.
- Use ONLY entities or terms that appear VERBATIM in the Paper context below.
- NEVER invent numbers, hardware brands, dataset names, affiliations, or citations.
- If the context does not contain enough details, OMIT them instead of commenting
    on missing information.
- Avoid bullet lists; write smooth narrative prose.
```

B. Persona-adapted examples for the Splitter

Below there are two real outline examples generated from the well-known paper *Attention Is All You Need* [42].

Example: Output for the General Public persona. The Splitter prioritises accessibility, conceptual clarity, and introductory explanations. The resulting outline highlights motivations and simplified descriptions of the model components.

```
- title: Introduction to Transformers
  description: Introduces the Transformer model, explaining why traditional RNNs
    are
    limited.
- title: Why Attention Matters
  description: Explains how attention mechanisms improve sequence transduction
    models
    compared to RNNs and CNNs.
- title: Model Architecture Overview
  description: Describes the basic components of the Transformer model, focusing on
    self-attention and feed-forward networks.
- title: Experimental Results
  description: Details the performance of the Transformer model on translation
    tasks,
    showing improved BLEU scores.
- title: Positional Encoding Explanation
  description: Describes how positional encodings help the Transformer model
    understand
    sequence order.
```

Example: Output for the Policy Maker persona. For an audience of policy makers, the Splitter shifts the emphasis towards societal implications, governance aspects, and impact, while still preserving the essential technical framing of the paper.

```
- title: Introduction to Transformers
  description: Introduces the limitations of recurrent models and proposes the
    Transformer
    architecture.
- title: Model Architecture Overview
  description: Describes the basic components of the Transformer model, including
    self-attention and feed-forward networks.
- title: Ethical Implications and Governance
  description: Discusses potential ethical concerns and governance issues related
    to the widespread adoption of Transformer models.
- title: Experimental Results and Performance
  description: Details the experimental setup and results of the Transformer model
    on machine translation tasks.
- title: Future Directions and Applications
  description: Outlines future research directions and potential applications of
    the
    Transformer model beyond translation.
```

C. Persona-adapted examples for the Storyteller

The following example is based on the paper *Attention Is All You Need* [42] and is included solely to illustrate how the Storyteller adapts tone and depth when the same content is generated for different personas. The same segment, *Model Architecture Overview*, is shown below for two audiences.

General Public.
Transformers revolutionized machine translation by simplifying the architecture and removing the need for recurrent networks. Instead of processing words one at a time, the model uses attention mechanisms to look at all parts of a sentence simultaneously, allowing for faster and more accurate understanding. This parallel processing makes training and inference significantly more efficient.

Policy Maker.
Transformer models eliminate both recurrent and convolutional layers, relying solely on self-attention mechanisms to process sequences. The encoder maps the input into continuous vector representations, which the decoder then uses to generate outputs through attention-based interactions. This architecture underpins many state-of-the-art systems and raises governance considerations regarding transparency and model interpretability.