Hallucination or Creativity: How to Evaluate AI-Generated Scientific Stories?

Alex Argese[1], Pasquale Lisena[1,*] and Raphaël Troncy[1]

[1]EURECOM, Sophia Antipolis, France

Abstract

Generative AI can turn scientific articles into narratives for diverse audiences, but evaluating these stories remains challenging. Storytelling demands abstraction, simplification, and pedagogical creativity – qualities that are not often well-captured by standard summarization metrics. Factual hallucinations are critical in scientific contexts, yet, detectors often misclassify legitimate narrative reformulations. In this work, we propose StoryScore, a composite metric for evaluating AI-generated scientific stories. StoryScore integrates semantic alignment, lexical grounding, narrative control, structural fidelity, redundancy avoidance, and entity-level hallucination detection into a unified framework. Our analysis also reveals why many hallucination detection methods fail to distinguish pedagogical creativity from factual errors, highlighting a key limitation: while automatic metrics can effectively assess semantic similarity with original content, they struggle to evaluate how it is narrated and controlled.

Keywords

Scientific Storytelling, Generative AI, AI evaluation, LLM, Hallucination Detection

1. Introduction

Generative AI and Large language models (LLMs) can summarize and rephrase complex content, presenting it in a narrative form that makes it accessible to non-experts [1]. Despite these advances, LLMs frequently struggle to appropriately adapt tone, level of detail, and stylistic choices to diverse audiences and communication objectives. In addition, LLMs are prone to hallucinations, eloquently stating unsupported claims, which pose a significant risk in scientific communication [2].

Recent works are exploring *scientific storytelling* as a way to generate structured and engaging explanation of scientific content and papers, tailored to specific personas [3, 4, 5]. Evaluating such narratives presents challenges fundamentally different from those faced in traditional summarisation tasks. Storytelling deliberately introduces abstraction, metaphor, and contextualization to enhance accessibility. At the same time, this complicates evaluation: assessing factual grounding becomes non-trivial when the generated text is not expected to closely mirror the source document. So, the question of how to evaluate generated narratives remains largely open. In particular:

(RQ1) How can the quality of a scientific story be evaluated beyond surface-level similarity to the source paper, while accounting for narrative coherence and persona adaptation?

(RQ2) How can hallucinations be reliably identified in stories where creative reformulation is not only expected but encouraged?

(RQ3) To what extent do existing hallucination detection methods conflate legitimate abstraction with factual inconsistency in scientific storytelling?

In this work, we address these questions in the context of designing a system that transforms scientific papers into audience-adapted stories. We make two main contributions:

1. *StoryScore*, a composite metric for automatic evaluation of scientific story generation;

R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): Proceedings of the Text2Story'26 Workshop, Delft (The Netherlands), 29-March-2026

*Corresponding author.

alex.argese@eurecom.fr (A. Argese); pasquale.lisena@eurecom.fr (P. Lisena); raphael.troncy@eurecom.fr (R. Troncy)

0009-0005-6151-5723 (A. Argese); 0000-0003-3094-5585 (P. Lisena); 0000-0003-0457-1436 (R. Troncy)

2. an empirical analysis of hallucination detection methods in persona-adaptive storytelling, showing why many approaches conflate creativity with hallucination.

The metrics are grounded in a concrete use case: a two-stage pipeline (a *Splitter* defining the outline and a *Storyteller* in charge of the writing) that transforms scientific papers into persona-adapted stories. The code used in this work is available at https://github.com/AlexArgese/ai-scientist-storyteller.

2. Related Work

Narrative Storytelling in Science Communication. Research shows that framing scientific content as a narrative significantly improves public comprehension, engagement [6] and recall of facts [7]. Scientific communication research emphasizes that effective dissemination requires deliberate choices about framing, analogy, structure, and level of detail, depending on the reader's background and goals [8, 9]. Recent work has also explored persona-driven or audience-aware generation to adapt explanations, tone, and terminology to reader profiles, showing benefits for accessibility and engagement but raising new requirements for evaluation beyond factuality alone [3, 4].

In human–AI interaction settings, researchers are exploring tools to help scientists crafting such narratives. A recent example is *RevTogether* that describes a multi-agent AI system that helps writers to revise "science stories" by blending engaging narrative structure with clear scientific content [10].

Automatic evaluation of generated narratives. Evaluating the quality of AI-generated summaries has also advanced beyond simple word-overlap scores. Traditional metrics like ROUGE [11] measure n-gram overlap with reference summaries, but they correlate poorly with human judgments [12, 13, 14]. Semantic metrics such as BERTScore [15] and MoverScore [16] use contextual embeddings to better capture semantic alignment and compare meaning rather than surface words. They achieve high correlation with human ratings on summarization, but remain largely insensitive to structural degradation, redundancy, and discourse-level control failures. This gap is amplified in multi-section stories, where discourse-level issues (e.g., global coherence or instruction leakage) strongly affect perceived quality while remaining weakly captured by similarity-based metrics. Other methods, such as QuestEval [13] and QAGS [17], ask questions about a summary and its source to detect inconsistencies.

Hallucination detection. Hallucination detection has been studied extensively in summarization and question answering, using entity-based checks, retrieval and entailment methods, and LLM-based judges [2, 18, 19]. Systems can use RAG-based techniques on specific documents or on the web, like the widely adopted GPTZero[1]. Recent empirical work shows pitfalls in automated detection. For example, they show that ROUGE and similar metrics can be easily fooled – e.g. repetitively duplicating correct content (adding length) artificially inflates ROUGE scores, even though no new facts are added [20]. In addition, common approaches often assume strict fidelity to the source. In persona-adaptive storytelling, legitimate reformulations and contextual expansions are expected. Thus, detectors can over-penalize creativity or behave unstably. We will critically analyse existing methods in Section 4.

3. The *StoryScore* metric for Generated Story Evaluation

The evaluation of stories generated from research papers, and in general, generated from a source material, requires metrics that balance faithfulness, completeness, and narrative quality under conditions of intentional abstraction. An effective evaluation framework should ensure that a generated story (i) **maintains semantic fidelity and representativeness** with respect to the source work, capturing the key content accurately even under extensive paraphrasing or reorganization; (ii) **preserves textual integrity**, minimizing artifacts and avoiding hallucinations introduced during generation; and (iii) **achieves structural and communicative adequacy**, ensuring coherent organization, appropriate

[1] https://gptzero.me/

Table 1
Summary of evaluation metrics used in this work. All components are in the $[0, 1]$ range.

Metric	Type	Objective
Context Recall	Lexical	Proportion of tokens from the original paper covered by the story
BERTScore	Semantic	Faithfulness of the story to the original content
Prompt Cleanliness	Structural	Absence of instruction leakage and prompt-related artifacts
Title Coverage	Structural	Similarity between generated and original section titles
NoRedundancy	Fluency	Avoidance of repeated n-gram loops and redundant phrasing patterns
NoHallucination	Factuality	Entity consistency (PERSON/ORG as detected by SpaCy) with the paper

titling, and elimination of unnecessary repetition. These requirements call for an evaluation approach that integrates complementary signals rather than relying on a single notion of similarity or correctness.

For this reason, we propose ***StoryScore***, a composite metric aggregating semantic similarity, lexical grounding, structural fidelity, fluency, and hallucination control into a single score in the range $[0, 1]$. Following extensive experimentation on a real use case, it is defined as:

$$\begin{aligned} \mathit{StoryScore} = (0.3\ \mathit{ContextRecall}) \ + \ (0.2\ \mathit{BERTScore}) \ + \ (0.2\ \mathit{PromptCleanliness}) + \\ (0.1\ \mathit{TitleCoverage}) \ + \ (0.1\ \mathit{NoRedundancy}) \ + \ (0.1\ \mathit{NoHallucination}) \end{aligned} \tag{1}$$

This formulation captures the most stable and reliable indicators, maintaining a balanced representation of narrative quality dimensions. The weights were selected heuristically after exploratory adjustments guided by manual qualitative inspection, with the goal of producing an aggregate score. The components of *StoryScore* are summarised in Table 1.

Context recall or **Article coverage** quantifies the amount of content from the original article that is reflected in the generated story, measured as the proportion of word-level lexical tokens from the article that also appear in the narrative (article-centered coverage). Tokens are defined as lowercased words, excluding punctuation and stopwords. Unlike metrics that rely solely on lexical overlap, contextual recall serves as a proxy for content coverage. Formally, it is defined as:

$$\mathit{ContextRecall} = \frac{|T_{\text{story}} \cap T_{\text{paper}}|}{|T_{\text{paper}}|} \tag{2}$$

where T_{paper} and T_{story} denote the sets of tokens extracted from the paper and the generated story. Higher values suggest a stronger connection to the original source material but also a vocabulary more faithful to the paper's language and less suited for a less knowledgeable audience. We deliberately adopt a simple set-based formulation to achieve a transparent and stable signal. This choice prioritises lexical grounding over semantic abstraction and may therefore disadvantage aggressive simplification.

BERTScore [15] measures **semantic faithfulness** by comparing contextual embeddings of the generated story with those of the source article. BERTScore aligns tokens from the hypothesis to the most semantically similar tokens in the reference using cosine similarity in the embedding space. For each token, the F1-score is computed by aggregating the highest-scoring matches. Formally, let H and R be the embedding sets for the hypothesis (story) and reference (paper), BERTScore is defined as:

$$\mathit{BERTScore} = F_1\big(\mathit{sim}(H, R),\ \mathit{sim}(R, H)\big) \tag{3}$$

where $\mathit{sim}(A, B)$ denotes the cosine-similarity-based alignment between tokens in A and their best matches in B. Higher values indicate stronger semantic grounding and better preservation of meaning.

Prompt Cleanliness measures the extent to which the generated story is free of **prompt-related artifacts** and **instruction leaks**, such as residual system directives, role indicators, formatting constraints, or meta-level indications. These artifacts indicate a failure of narrative abstraction, where the model collapses from storytelling to instruction-following behaviour, severely degrading readability. The generated text is analysed line by line and sentence by sentence using a set of regular expression

patterns that we empirically found as relevant. A contamination score C is computed as:

$$C = \frac{1.0 \cdot N_{\text{line}} + 0.75 \cdot N_{\text{sent}} + 1.25 \cdot N_{\text{json}} + 0.75 \cdot N_{\text{fence}} + 2.5 \cdot N_{\text{block}}}{|L|} \tag{4}$$

where L denotes the set of non-empty lines in the generated text, and:

- N_{line} is the number of lines beginning with explicit instruction markers (e.g., `Human:`, `Rules:`);
- N_{sent} counts sentences that exhibit imperative instruction patterns at sentence boundaries;
- N_{json} denotes the number of lines consisting solely of structured JSON-like content;
- N_{fence} counts occurrences of markdown code fences;
- N_{block} counts dense instruction blocks characterized by multiple repeated occurrences of imperative constraints (e.g., three or more instances of "do not" within a single sentence or paragraph).

The weights associated with each term are selected empirically to prioritize the detection of severe leakage patterns over recall, assigning stronger penalties to artifacts that indicate a near-complete collapse into instruction-following mode. The score is clipped to the unit, as outputs exceeding this threshold are considered equally dominated by prompt artifacts. Prompt Cleanliness is then defined as:

$$\mathit{PromptCleanliness} = 1 - \min(1, C) \tag{5}$$

Title Coverage evaluates whether the generated story preserves the **section structure** of the target outline[2]. Let $\mathcal{O} = \{o_1, \ldots, o_5\}$ denote the target section titles and $\mathcal{G} = \{g_1, \ldots, g_5\}$ the generated ones. We compute a soft similarity score and average it across sections:

$$\mathit{TitleCoverage} = \frac{1}{n} \sum_{i=1}^{n} \text{sim}\big(\text{norm}(g_i),\ \text{norm}(o_i)\big) \tag{6}$$

where n is the number of sections, $\text{norm}(\cdot)$ removes differences in case, whitespaces, and punctuation, and $\text{sim}(\cdot,\cdot) \in [0, 1]$ is a string similarity function (1 for identical titles, lower values for partial matches). This yields a graded measure of structural fidelity that is robust to minor formatting differences.

No Redundancy is a **fluency indicator** that penalizes degenerative loops and excessive reuse of the same textual fragments, which are common artifacts in long-form generation. Repetition is quantified by computing the frequency of word-level n-grams in the generated story, with $n = 3$ (trigrams), which provide a robust balance between sensitivity to repetition and tolerance to natural phrasing. Let $\mathcal{G}_n$ denote the multi-set of all n-grams extracted from the narrative, the repetition rate is defined as:

$$\mathit{RedundancyRate} = \frac{\max_{g \in \mathcal{G}_n} \text{freq}(g)}{|\mathcal{G}_n|}, \qquad \mathit{NoRedundancy} = 1 - \mathit{RedundancyRate} \tag{7}$$

High values of *NoRedundancy* indicate fluent, varied text with minimal looping or redundant phrasing. This formulation is designed to capture obvious degenerative loops rather than fine-grained stylistic repetition, prioritizing robustness and interpretability over sensitivity to subtle discourse patterns.

No Hallucination quantifies the inclusion of entities that are not supported by the source material. After a post-processing normalization step applied to extracted entities (e.g. lowercasing, removal of possessive suffixes), we perform a NER-based comparison: the generated story is analyzed using SpaCy [21] to extract entities of type PERSON and ORG, which are then matched against the entities detected in the source paper. The restriction to these entity types is deliberate: in our manual qualitative inspection of generated outputs, hallucinations most frequently manifested as invented author names and institutional affiliations, making PERSON/ORG the most relevant entity types for a stable automatic signal. Any entity appearing in the story but absent from the source material is treated as a hallucination. A broader discussion of alternative hallucination detectors is provided in Section 4. Formally, let *GeneratedEntities* be the set of PERSON/ORG entities extracted from the story, and *HallucinatedEntities* the subset of those not found in the paper, the score is defined as:

$$\mathit{NoHallucination} = 1 - \frac{|\mathit{HallucinatedEntities}|}{|\mathit{GeneratedEntities}|} \tag{8}$$

[2]In our use case, this is the outline generated by the *Splitter* module, as in Section 1.

4. Hallucination Detection in Scientific Storytelling

Unlike summarization, storytelling intentionally involves simplification, contextualization, and narrative adaptation, producing an expected and desirable creative divergence. The core challenge is in distinguishing *legitimate narrative abstraction* from *true factual hallucination*, a distinction that existing hallucination detection methods (largely designed for summarization) are not well equipped to make.

The following analysis is grounded in a qualitative inspection of AI-generated stories produced by the pipeline on a representative subset of papers from the corpus. Each hallucination detection method was applied to the same generated stories and its outputs were manually examined to identify systematic failure modes specific to persona-adaptive scientific storytelling.

In this setting, hallucination cannot be reduced to mere deviation from the source text: a story may remain faithful while employing metaphors, analogies, or contextualization absent from the original paper. This inherent ambiguity makes hallucination detection particularly challenging and motivates the comparison of multiple detection approaches.

Capitalised Words as Entity Proxies. This approach is deliberately simplistic: any word starting with a capital letter is treated as a potential factual entity. If such a token appears in the story but not in the article, it is marked as hallucination. This heuristic reveals clear limitations:

- abbreviations in the story (e.g. "AI") are flagged if the source document only contained the expanded form (e.g. "Artificial Intelligence"),
- metaphors or narrative constructs capitalised for emphasis are misclassified as entities,
- minor morphological variants (pluralisation, genitives) produce false positives.

Although rudimentary, this method lays the foundation for understanding the structure of the problem: hallucination detection must separate surface-form noise from true semantic divergence.

NER-Based Detection (SpaCy PERSON/ORG). We leverage SpaCy to only detect named entities of types PERSON and ORG, as described in Section 3. This significantly reduces the number of candidates and detects genuinely invented organisations or people introduced by the model. Nevertheless, NER remained a shallow signal: it can detect incorrect affiliations, but fails to identify deeper factual inconsistencies, e.g. wrong scientific claims, invented datasets, or unsupported causal statements. We specifically focus on PERSON/ORG because our qualitative analysis repeatedly found fabricated authors and affiliations to be the most frequent and disruptive hallucination pattern in our stories.

MIRAGE Rewrite-Consistency Scoring. MIRAGE is a library for hallucination detection based on rewriting the same passage multiple times and measuring the stability of the appearing concepts [22]. If an idea disappears or mutates across rewrites, MIRAGE treats it as hallucinated. Although effective in summarization, this approach does not transfer well to persona-oriented storytelling. In our experiments, MIRAGE consistently and incorrectly flags explanatory metaphors (e.g., introducing a system by analogy with a familiar real-world process that is not explicitly described in the paper) and rephrasing for non-expert audiences (e.g., replacing technical terminology with higher-level conceptual descriptions). This indicates a misalignment between MIRAGE's operational definition of fidelity, centred on literal grounding, and the goals of narrative systems that intentionally simplify the source material.

LLM-as-a-Judge. We use an LLM to judge hallucinations directly. The judge receives the full scientific article as `CONTEXT`, the story as `ANSWER`, and produces a structured JSON verdict that includes faithfulness, hallucinated entities and numerical errors. This brings a qualitative leap because the LLM can reason about paraphrases and understand the broader context.

A first experiment conducted with the Qwen2.5-7B model [23] has shown two systematic patterns:

1. **Missed hallucinations:** Some fabricated facts (e.g. invented affiliations) are left undetected.

Table 2
Comparison of explored hallucination detection methods.

Method	What it detects	Strengths	Weaknesses
Capitalised Words	Surface-form mismatch	Simple, transparent	Noisy; overflags creativity
SpaCy NER	Incorrect entity mentions	Good precision	Misses conceptual hallucinations
MIRAGE	Rewrite instability	Captures semantic drift	Penalises analogies
LLM-as-Judge (Qwen)	Factual consistency	Understands paraphrases	Inconsistent; invents hallucinations
LLM-as-Judge (GPT 5.1)	High-level reasoning errors	Close to human judgement	Too strict for storytelling
HHD	Entity + retrieval alignment	Balanced approach	Unstable thresholds; mixed reliability

2. **False positive:** In several cases, it labeled hallucinated entities that were explicitly supported by the source paper (and mentioned in the story), i.e., the judge itself hallucinated the hallucination.

In short, the judge sometimes *hallucinates hallucinations*, making it too unreliable.

A second experiment used the GPT 5.1 model [24]. While its assessments were more consistent with human judgment, the model remained overly strict, labelling benign contextual expansions as hallucinations merely because the source material did not explicitly mention those cases.

Hybrid Hallucination Detection (HHD). A collection of some positive results from previous discoveries are combined together to form a hybrid technique consisting to:

1. extract "technical tokens" via SpaCy (capitalised words, acronyms, numbers),
2. retrieve the top-k most similar sentences from the article using MiniLM embeddings [25],
3. mark a token as hallucinated only if it appears in none of the retrieved contexts *and* the similarity score is below a threshold.

This approach combines symbolic robustness (entity extraction), semantic flexibility (retrieval-based context), and local grounding (sentence-level comparison), but is difficult to calibrate.

Manual inspection on generated stories showed that false positives dominate the detector's output, mainly due to pedagogical reformulations that are conceptually faithful but not literally supported by retrieved sentences. For example, the following excerpt is from a generated story about [26]:

> ***Hermes**, the messenger god of ancient Greece, was known for his speed and efficiency. Similarly, the **HERMES** system acts as a swift messenger between the initial prompt and the final, refined medical image segmentation.*

In this case, *HERMES* is the name of a framework, introduced through a creative but semantically correct analogy, yet incorrectly flagged as hallucinated. Conversely, false negatives occurred when the retrieval step returned semantically adjacent but non-supportive contexts. Another example from the same story fails to correctly assign the acronym *FM* to *Foundation Models*, resulting in the following undetected hallucination:

> *The proposed solution is an automated framework designed to enhance the accuracy of **flash memory (FM)-based segmentation**.*

Because these two error types move in opposite directions, tuning the threshold did not lead to a stable operating point, making HHD unsuitable as a component for StoryScore.

Table 2 summarizes our findings on different hallucination detection techniques, that lead to two key outcomes. First, The most operationally stable detector was the simplest one, namely the NER-Based

Table 3
StoryScore components on the test set, comparing the pre-trained and fine-tuned versions of the pipeline.

LLM version	**StoryScore**	BERTScore	Context Recall	Prompt Cleanliness	Title Coverage	No Redundancy	No Hallucination
Pre-trained	0.560	0.780	0.390	0.011	0.990	0.893	0.957
Fine-tuned	0.787	0.815	0.472	1.000	0.998	0.903	0.925

detection, combined with regex normalisation. In particular, it was the only method that: 1. remained stable across papers, showing consistent behaviour across different documents, 2. did not penalise metaphors, 3. detected genuine fabricated entities, and 4. integrated cleanly with a software pipeline, without the need of external and costly API calls. Second, being either too naive (NER), too strict (MIRAGE), or too unstable (LLM judge), the hallucination metric must have a reduced weight in the overall metric (limited to 10%) to inform without overshadowing the other reliable qualities.

5. Preliminary Findings

We apply StoryScore to a set of 76 stories generated by our use case pipeline. Table 3 reports aggregated statistics comparing two versions of the pipeline, using either a pre-trained version of Qwen2.5, or a fine-tuned version of the model. Fine-tuning substantially improves StoryScore and completely eliminates prompt leakage. Prompt Cleanliness and Title Coverage saturated, but their presence in StoryScore is a diagnostic safeguards. Nevertheless, the pre-trained model still attains positive scores, particularly on BERTScore, Title Coverage, No Redundancy, and No Hallucination.

A manual inspection found that narratives generated in the pre-trained settings are consistently affected by prompt leakage, redundancy, and excessive narrative filler, deficiencies that are only weakly penalized by global semantic metrics. The latters are tolerant of paraphrasing, repetition, and generic formulations, rewarding long and semantically "safe" texts despite being poorly readable. Conversely, the fine-tuned model generates more controlled, dense, and readable narratives, yet these qualitative improvements are only partially reflected in metrics. Overall, variations in StoryScore were consistent with qualitative judgments of readability, narrative control, and factual grounding.

6. Conclusions and Future Work

We introduced StoryScore, a composite metric that integrates semantic alignment, lexical grounding, structural fidelity, fluency, and hallucination control, and analysed its behaviour on a concrete scientific storytelling use case. The combination of complementary metrics provides a more informative picture of system behaviour than any single score alone. The initial evaluation suggests that StoryScore is useful for comparisons between generated stories, even if insufficient for discriminating critical aspects of narrative quality and text control in an absolute way.

Hallucination detection proved to be an intrinsically complex task, even more when creativity and pedagogical reformulation are involved. Existing hallucination detection approaches are either too shallow, too rigid, or too unstable. Future evaluation frameworks should explicitly account for persona adaptation and narrative transformation, calling for a new definition of hallucination that distinguish between acceptable abstraction and factual distortion. Finally, this work motivates further refinement of composite metrics such as StoryScore – in particular to address narrative control and structural validity – as well as an assessment of alignment with human judgement following some best practices [14].

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 authors used GPT5.2 for grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] 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: C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2025, Association for Computational Linguistics, Suzhou, China, 2025, pp. 13954–13966. doi:10.18653/v1/2025.findings-emnlp.750.

[2] 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 Comput. Surv. 55 (2023). doi:10.1145/3571730.

[3] J. Roschelle, M. Bansal, G. Biswas, C. Hmelo-Silver, J. Lester, Generative AI Unlocks the Power of Interactive Storytelling for Science Teachers and Learners, Social Innovations Journal 30 (2025). URL: https://socialinnovationsjournal.com/index.php/sij/article/view/9993.

[4] Y. Li, Y. Wang, Y. Lee, H. Chen, A. N. Petri, T. Cha, Teaching Data Science through Storytelling: Improving Undergraduate Data Literacy, Thinking Skills and Creativity 48 (2023) 101311. doi:10.1016/j.tsc.2023.101311.

[5] 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.

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

[7] 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.

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

[9] 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.

[10] Y. Zhang, K. Fu, Z. Lu, RevTogether: Supporting Science Story Revision with Multiple AI Agents, in: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA '25, Association for Computing Machinery, New York, NY, USA, 2025, pp. 1–7. doi:10.1145/3706599.3719888.

[11] 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/.

[12] 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 - Volume 3: ICAART, INSTICC, SciTePress, 2023, pp. 682–689. doi:10.5220/0011744500003393.

[13] T. Scialom, P.-A. Dray, S. Lamprier, B. Piwowarski, J. Staiano, A. Wang, P. Gallinari, QuestEval: Summarization Asks for Fact-based Evaluation, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 6594–6604. doi:10.18653/v1/2021.emnlp-main.529.

[14] C. Chhun, P. Colombo, F. M. Suchanek, C. Clavel, Of Human Criteria and Automatic Metrics: A Benchmark of the Evaluation of Story Generation, in: 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 5794–5836. URL: https://aclanthology.org/2022.coling-1.509/.

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

[16] 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: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), 2019 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.

[17] A. Wang, K. Cho, M. Lewis, Asking and Answering Questions to Evaluate the Factual Consistency of Summaries, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5008–5020. doi:10.18653/v1/2020.acl-main.450.

[18] 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 Trans. Inf. Syst. 43 (2025). doi:10.1145/3703155.

[19] 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: 2025 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.

[20] 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: C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng (Eds.), 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2025, pp. 34728–34745. doi:10.18653/v1/2025.emnlp-main.1761.

[21] I. Montani, M. Honnibal, A. Boyd, S. V. Landeghem, H. Peters, explosion/spacy: v3.7.2: Fixes for apis and requirements, 2023. doi:10.5281/zenodo.10009823.

[22] 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, CIKM '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 6539–6543. doi:10.1145/3746252.3761644.

[23] 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.

[24] OpenAI, ChatGPT 5.1, https://chat.openai.com, 2025. Large Language Model.

[25] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, MINILM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers, in: 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020, pp. 5776–5788.

[26] Y. Gao, Training like a medical resident: Context-prior learning toward universal medical image segmentation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 11194–11204.